UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA ANIMAL

An evolutionary perspective on the circadian system in western Iberian *Squalius* freshwater species

João Miguel Moreno

Mestrado em Biologia Evolutiva e do Desenvolvimento

Dissertação orientada por:
Professora Doutora Maria Manuela Coelho

2018

Acknowledgements

Agradecimentos

Em primeiro lugar gostava de agradecer a todas as pessoas que, durante todo o meu percurso académico, contribuíram de alguma forma para o meu crescimento científico e, que de forma indireta, permitiram que eu chegasse até aqui.

Passo a agradecer, agora, aos meus colegas do grupo *Evolutionary genetics* do cE3c que, durante o projeto, contribuíram ativamente na discussão de ideias e que me fizeram companhia durante os períodos de trabalho na sala 12 e horas de almoço. Bora lá, amigos *Squalius* (e cetáceos)!

Agradeço também ao Tiago Jesus e ao David que ajudaram a rever o trabalho final.

Queria agradecer aos meus amigos por todo o apoio que me deram durante todo o meu percurso académico. No entanto, gostava de fazer um agradecimento extra a uma pessoa especial (e sim, tu sabes que me refiro a ti) por me ajudares a ultrapassar os piores momentos de todo este percurso. Foi também graças a ti que hoje atinjo este grande objetivo da minha vida. Este trabalho também é para ti!

Um enorme agradecimento à minha família (aos que estão e aos que já não estão entre nós), por me proporcionarem a oportunidade de poder traçar este caminho académico que escolhi e por terem sempre acreditado em mim, seja enquanto pessoa e/ou investigador. Estarão sempre comigo e este trabalho é também dedicado a vocês!

Por fim, faço um agradecimento especial à minha orientadora, a (grande) Professora Manuela Coelho, que sempre me ajudou em todos os momentos - bons e maus -, que sempre acreditou no meu potencial e, acima de tudo, esteve, em todos os momentos, disponível para ouvir as minhas ideias, que me deixou integrar duas áreas de investigação de que tanto gosto, que me estimulou a aprender mais e me ensinou grande parte do que hoje sei de Evolução (e de *Squalius*!). Jamais esquecerei todas as oportunidades que me proporcionou, mas também toda a paciência que teve comigo! Um grande obrigado.

Abstract

The circadian system is a biological timing system that improves the inherent ability of organisms to deal with environmental instabilities, creating oscillations that are highly synchronised with daily fluctuations of light in several biochemical processes. Temperature has been pointed as an integrant agent in several aspects of the circadian system. Fish, as ectothermic organisms, have an increased dependence on the environmental temperature to maintain and optimize the circadian system. Here we study the freshwater fish *Squalius* genus, which is represented in Portugal by four species distributed across a latitudinal gradient with variable environmental conditions of light and temperature. *Squalius carolitertii* inhabits the Atlantic-type northern basins, while the sister species *S. torgalensis* and *S. aradensis* inhabit the Mediterranean-type southern basins. Additionally, *S. pyrenaicus* has a broader distribution, inhabiting both Atlantic-type northern basins (e.g. Tagus) and Mediterranean-type southern basins (e.g. Almargem). This distribution gives the opportunity to study species adaptation to different environmental backgrounds. Analysis of *S. torgalensis* and *S. carolitertii* transcriptomes and comparison with light-induced zebrafish transcriptomes revealed four main gene families involved (Cryptochromes, Period, CLOCK and BMAL) in the circadian system, and one single-copy gene (*timeless*) whose function was possibly related to the circadian system. The cDNA of these identified genes was re-sequenced in *S. torgalensis* and *S. carolitertii* and sequenced *de novo* for *S. pyrenaicus* and *S. aradensis*. Characterisation of protein-protein interactions revealed that the studied proteins retain functions in the circadian system common to other vertebrates, but also display signals of diversification. Protein-protein interactions for TIMELESS revealed this protein presents a differentiation in its function towards the cell cycle regulation but retains the ability to act as a circadian regulator. Results of phylogenetic analysis of each family, based on predicted protein sequences for *Squalius* species, were, for some proteins (CRY1BA, PER1A, PER2, BMAL2 and TIMELESS), congruent with the phylogenies of other nuclear genes from these species. Moreover, by using an integrative approach involving tests to detect signatures of selection and a structural and functional protein analysis, it was possible to infer N-S patterns consistent with adaptation in these species, probably related with the different environmental conditions they experience. Additionally, *S. torgalensis* seems to present a pattern of local adaptation in two proteins (CRY1AA and PER1A) that could be explained by the specific environmental conditions experienced in Mira basin. Lastly, a well-supported convergence for some proteins (CRY1BB, CRY3, PER3, CLOCKA, CLOCKB, BMAL1A and BMAL1B) was found in southern populations of *S. aradensis* and *S. pyrenaicus* from Almargem. Signatures of selection supported this convergence in *clocka*, but for other genes, convergence was only exposed after protein characterisation. Convergence was evident because these proteins display similar features, specifically in physicochemical parameters and post-translational modifications patterns, congruent with the more similar environmental conditions they experience. Also, our results support that the evolution of circadian-related proteins in these species has been strongly driven by temperature, as thermostability was found to be the protein feature presenting most modifications in all protein.

Keywords: Iberian freshwater fish; circadian rhythms; adaptation; dN/dS; protein characterisation

Resumo

O ritmo circadiano constitui um importante sistema biológico que permite a sincronização de diversas funções biológicas dos organismos de acordo com as flutuações diárias de luz. Sendo assim, atua como um relógio endógeno por um período de ~24h, sendo regulado a nível molecular por redes de retroação compostas por elementos positivos e negativos que, respetivamente, ativam ou reprimem a expressão de genes alvo. Estas redes são mantidas por quatro famílias de genes (CRY, PER, CLOCK e BMAL). As proteínas CLOCK e BMAL atuam como um heterodímero (CLOCK:BMAL), ligando-se às *E-boxes* no ADN de forma a ativar a expressão dos genes alvo, incluindo os elementos negativos do ciclo, os genes *cry* e *per*. Após a tradução dos genes referidos, as proteínas CRY e PER formam também um heterodímero (CRY:PER), que interage com o heterodímero CLOCK:BMAL, levando à repressão da expressão génica, incluindo os próprios *cry* e *per*. O ciclo repete-se quando os níveis de CRY:PER são baixos o suficiente para o CLOCK:BMAL poder novamente ligar-se ao ADN e estimular a expressão génica. Adicionalmente, o gene *timeless,* que codifica para a proteína TIMELESS, mostrou ser uma parte central do ritmo circadiano em *Drosophila,* mas as suas funções nos vertebrados permanecem ainda pouco estudadas, tendo sido proposto que a proteína TIMELESS possa desempenhar um papel secundário no ritmo circadiano, ao servir de alternativa às proteínas PER. As consequências das interações moleculares que ocorrem no ritmo circadiano manifestam-se, entre outras, na otimização do metabolismo, regulação da expressão génica, e resposta a alterações ecológicas e ambientais. Todo o sistema do ritmo circadiano encontra-se, ele próprio, sincronizado pelas flutuações diárias de luz. Contudo, foi também proposto que a temperatura ambiental apresenta um papel fulcral na sincronização e manutenção do ritmo circadiano.

As famílias de genes envolvidas no ritmo circadiano têm vindo a ser extensivamente caracterizadas, em particular nos peixes, organismos que possuem um acrescido número de parálogos face a outros vertebrados. Para além disto, os peixes possuem diversas adaptações a nível do ritmo circadiano e temperatura, pois sendo ectotérmicos, apresentam uma maior dependência face à temperatura ambiental para desempenhar as suas funções fisiológicas.

O género *Squalius* encontra-se representado nos rios Portugueses por quatro espécies (*S. carolitertii, S. pyrenaicus, S. torgalensis* e *S. aradensis*), distribuídas ao longo de um gradiente latitudinal de luz e temperatura. *S. carolitertii* é a espécie que habita os rios da região a norte do Tejo (ex: Mondego), *S. pyrenaicus* apresenta uma distribuição mais alargada, habitando as bacias do centro e sul (ex: Tejo e Almargem), enquanto as espécies *S. torgalensis* e *S. aradensis,* se encontram confinadas a pequenas bacias no sudoeste de Portugal (ex. Mira e Arade). A esta distribuição encontra-se subjacente uma variação ambiental latitudinal, uma vez que as bacias das regiões norte e centro estão sob influência Atlântica, apresentando temperaturas mais baixas, enquanto as bacias da região sul se encontram sob influência Mediterrânea, apresentando, de um modo geral, temperaturas mais altas e onde ocorrem períodos de seca durante o Verão. Existe também uma variação na duração e intensidade da luminosidade ao longo deste gradiente latitudinal. Atendendo ao mencionado, estas espécies apresentam-se como um modelo de particular interesse no estudo da evolução adaptativa, nomeadamente ao nível do ritmo circadiano e sua relação com a temperatura.

No presente estudo, foram identificados os genes envolvidos no ritmo circadiano dos peixes dulciaquícolas do género *Squalius* com base numa análise dos transcriptomas de *S. carolitertii* e *S. torgalensis* recentemente publicados, e sua comparação com resultados de outros transcriptomas de *Danio rerio* (aqui usado como organismo de referência), obtidos após a exposição a estímulos luminosos. Identificaram-se quinze genes pertencentes às quatro grandes famílias já descritas: seis genes da família CRY (*cry1aa, cry1ab, cry1ba, cry1bb, cry2* e *cry3*); quatro genes da família PER (*per1a, per1b, per2* e *per3*); dois genes da família CLOCK (*clocka* e *clockb*); e três genes da família

BMAL (*bmal1a, bmal1b* e *bmal2*). Adicionalmente, foi identificado o gene *timeless*. Com exceção da família CLOCK, na qual não foi possível identificar o gene *clock2*, a constituição de todas as famílias encontra-se de acordo com o descrito para outros peixes, nomeadamente *D. rerio.* Os genes referidos foram sequenciados para todas as espécies em estudo, através de ADN complementar, tendo sido as sequências destes posteriormente utilizadas para prever as sequências das proteínas para as quais codificam.

Através das sequências das proteínas foi possível caracterizar os padrões de interação proteína-proteína, que revelaram que estas proteínas retêm as suas funções ancestrais como reguladores circadianos, exibindo no entanto, sinais de diversificação pós-duplicação que se observam através de interação das proteínas CRY e PER com outras proteínas cuja função se encontra mais relacionada com a resposta à temperatura. Isto, não só mostra a sua diversificação como também reforça a relação entre o ritmo circadiano e a temperatura na otimização das oscilações circadianas. Para a proteína TIMELESS, esta análise permitiu clarificar a sua função nas espécies estudadas, já que, apesar da proteína manter a capacidade de interagir com as proteínas CRY e, assim, manter uma potencial função como regulador circadiano, a maior parte das suas interações dão-se com proteínas relacionadas com a regulação do ciclo celular.

A análise filogenética de cada família de genes, conseguida a partir das sequências das proteínas previstas, permitiu mostrar as relações evolutivas das duplicações nestas espécies do género *Squalius*, confirmando que a história evolutiva dos genes destas famílias é idêntica ao já descrito para outras espécies de peixes.

Adicionalmente, integrando testes estatísticos para determinar as assinaturas de seleção a nível molecular (dN/dS) com a caracterização estrutural e funcional das proteínas, foi possível também estudar a adaptação destas espécies às diferentes condições de luz e temperatura a que estão sujeitas. Para as proteínas CRY1BA, PER1A, PER2, BMAL2 e TIMELESS, o padrão encontrado parece refletir a história evolutiva das espécies do género *Squalius* encontrada nas filogenias de outros genes nucleares.

Foi também encontrada, para outras proteínas, uma diferenciação Norte-Sul consistente a variação de luz e temperatura a que espécies estão sujeitas nos seus habitats. Para além disso, verificou-se que em *S. torgalensis*, para as proteínas CRY1AA e PER1A, os sinais desta adaptação local são notórios nos resultados dos testes de seleção, que mostram que, para estas proteínas, *S. torgalensis* apresenta um forte sinal de seleção positiva. Além disso, a caracterização das proteínas, nomeadamente através dos parâmetros físico-químicos e padrões de modificações pós-tradução, mostra que esta espécie se encontra particularmente diferenciada face às restantes. As evidências de adaptação local podem ser explicadas pelas condições particulares de luz e temperatura no rio Mira, nomeadamente, o maior ensombramento causado pela vegetação ripária e a temperatura média da água elevada, mas mais estável do que nos restantes rios do sul.

Encontrou-se, também, uma convergência entre as populações de *S. aradensis* e *S. pyrenaicus* de Almargem, nas proteínas CRY1BB, CRY3, PER3, CLOCKA, CLOCKB, BMAL1A e BMAL1B. As assinaturas de seleção permitiram suportar esta convergência no gene *clocka*, mas, nos restantes genes, esta convergência tornou-se evidente apenas após a caracterização das proteínas, uma vez que, nas referidas espécies, estas proteínas apresentam características semelhantes, nomeadamente a nível dos parâmetros físico-químicos e modificações pós-tradução. Estas características das referidas proteínas, refletem-se em valores de maior termoestabilidade nestas espécies, muito provavelmente relacionadas com o facto destas espécies habitarem rios com valores muito próximos de luz e temperatura.

No geral, os resultados obtidos para a termoestabilidade da maior parte das proteínas estudadas, sugerem que a temperatura é um fator seletivo que se relaciona com a evolução destas proteínas nas espécies do género *Squalius* estudadas. Deste modo, também poderemos dizer que estas espécies constituem um bom modelo para a continuação destes estudos, nomeadamente ao explorar a relação entre o ritmo circadiano e a temperatura de forma mais detalhada.

Por fim, foi possível mostrar que a abordagem inovadora de integração dos resultados dos testes de seleção baseados no rácio dN/dS com a caracterização de proteínas agora apresentada, permite detetar sinais de seleção, cuja estatística associada a estes testes não permite, só por si, detetar.

Palavras-chave: Peixes de água doce Ibéricos; ritmo circadiano; adaptação; dN/dS; caracterização de proteínas

Table of Contents

List of Tables and Figures

List of Abbreviations

bHLH – basic helix-loop-helix

BLAST – Basic local alignment search tool

BMAL – Brain and Muscle ARNT (Aryl hydrocarbon Receptor Nuclear Translocator)-Like

cDNA – Complementary deoxyribonucleic acid.

CLOCK – Circadian locomotor output cycles protein kaput

CRY – Cryptochrome

DNA – Deoxyribonucleic acid

GO – Gene ontology

HSF – Heat shock factor

HSP – Heat shock protein

KW – Kruskal-Wallis

PAS - Period-Aryl hydrocarbon receptor nuclear translocator- Single mind domain

PCR – Polymerase chain reaction

pI – Isoelectric point

PER – Period

PPI – Protein-protein interaction

PTM – Post-translational modification

RNA – Ribonucleic acid

1. Introduction

Organisms are exposed to daily fluctuations in their natural environment. To overcome environmental fluctuations, organisms developed biological timing systems to optimize their physiological and biochemical processes in space and time (Foulkes et al. 2016). These systems work as internal clocks and require a proper synchronization with environmental signals that specify the actual time of the day.

The circadian system is a universal biological timing system found virtually in all organisms (Paranjpe & Sharma 2005). Circadian system is synchronised by light–dark cycle of a day's period and present oscillations with a period of ~24h called circadian rhythms. Oscillations are generated and regulated at molecular level, but the outcomes have been shown to influence several aspects of physiology, behaviour and ecology of organisms (Paranjpe & Sharma 2005; Vaze & Sharma 2013). In fact, circadian rhythms, when properly entrained by light, improve the inherent aptitude of organisms to survive under ever-changing environments by aiding them to efficiently anticipate periodic events, specifically light changes and climate seasons (Paranjpe & Sharma 2005; Vaze & Sharma 2013).

The molecular circadian system consists of a network of signalling transduction pathways regulated mainly by interconnected transcription-translation feedback loops (Figure 1.1) (Dunlap 1999; Pando & Sassone-Corsi 2002). The regulatory loops are sustained by the so-called core circadian genes and proteins and require about 24h to complete a cycle (Pando & Sassone-Corsi 2002; Foulkes et al. 2016). In vertebrates, several genes have been reported to be responsible for the maintenance and regulation of the circadian system (Foulkes et al. 2016). The core circadian-genes belong to four main gene families: Cryptochromes (CRY), Period (PER), CLOCK, and BMAL (Pando & Sassone-Corsi 2002). These gene families encompass several characterized genes (*cry, per, bmal* and *clock*) in vertebrates, namely in fish for which these families possess a larger number of circadian paralogs as compared to the other vertebrates (Toloza-Villalobos et al. 2015). In *Danio rerio,* there have been identified six *cry* genes (*cry1aa, cry1ab, cry1ba, cry1bb, cry2, cry3*), four *per* genes (*per1a, per1b, per2, per3*), three *bmal* genes (*bmal1a, bmal1b, bmal2*) and, three *clock* genes (*clocka, clockb, clock2*) (Wang 2008a, 2008b, 2009; Liu et al. 2015). Cryptochrome genes encode for a class of flavoproteins that are sensitive to blue light (Lin & Todo 2005) and the period genes encode for proteins that also display a strong but differential light responsiveness (Pando et al. 2001; Vatine et al. 2009; Vallone et al. 2004). They were found to be key agents in the entrainment of the circadian system, as they are photoreceptors responsible for transducing light signals to the core circadian machinery (Tamai et al. 2007). BMAL (Brain and muscle ARNT like) and CLOCK (Circadian locomotor output cycle kaput) families encode for canonical circadian proteins, a highly conserved bHLH (basic-Helix-Loop-Helix)-PAS (Period-Aryl hydrocarbon receptor nuclear translocator- Single mind) transcriptional factors and are the positive elements of the circadian system (Pando & Sassone-Corsi 2002; Foulkes et al. 2016).

In vertebrates, the whole network starts with the proteins CLOCK and BMAL. These act as heterodimers CLOCK:BMAL and bind to E-box enhancer elements in DNA to trigger transcription of target genes (Pando & Sassone-Corsi 2002; Foulkes et al. 2016). E-box enhancers are present in promoter regions of a vast spectrum of genes including the negative elements of the network, *per* and *cry* genes (Nakahata et al. 2008). After activation of *per* and *cry* transcription and subsequently translation, PER and CRY proteins are translocated to the nucleus in the form of an heterodimer

PER:CRY which interacts with the CLOCK:BMAL heterodimers to inhibit their function and thereby downregulating expression of E-box-dependent genes (Ishikawa et al. 2002; Pando & Sassone-Corsi 2002). Hereafter, there is a reduction in the expression of *per* and *cry*, and levels of the negative elements drop. When PER and CRY proteins are reduced to critical levels, CLOCK and BMAL become again able to bind E-box elements and activate expression of these genes and the cycle is repeated (Reppert & Weaver 2001; Pando & Sassone-Corsi 2002). This loop of activation/repression takes ~24h to complete one cycle and it has been described as the fundamental molecular source of circadian rhythmicity (Reppert & Weaver 2001; Pando & Sassone-Corsi 2002). In addition, *bmal* genes are regulated and rhythmically expressed under the control of another feedback loop involving two orphan nuclear receptors, REV-ERBα and RORα (Preitner et al. 2002). Even though REV-ERBα and RORα regulate *bmal* expression, they are not considered to belong to the core-circadian apparatus. Lastly, *timeless* gene was described as having a possible role on circadian regulation in vertebrates (Barnes et al. 2003; Gotter 2006; Yoshizawa-Sugata & Masai 2007). Although its function has been widely described for several invertebrates (e.g. *Drosophila melanogaster*), there is a lack of information concerning its role in the circadian system of vertebrates, including for fish (Gotter 2006). Several authors hypothesize *timeless* function may mainly connected with the CRY/PER feedback loop, with TIMELESS protein serving as potential alternative to the PER protein in the negative phase of the cycle, as it happens in *Drosophila* circadian system (Gotter 2006). To complete the circadian regulatory network, further layers of regulation are provided by post-transcriptional and post-translational modifications (McClung 2011; Lim & Allada 2013), epigenetic regulation, and metabolic networks (Asher & Sassone-Corsi 2015). They allow to reinforce the core feedback loops and to increase their robustness.

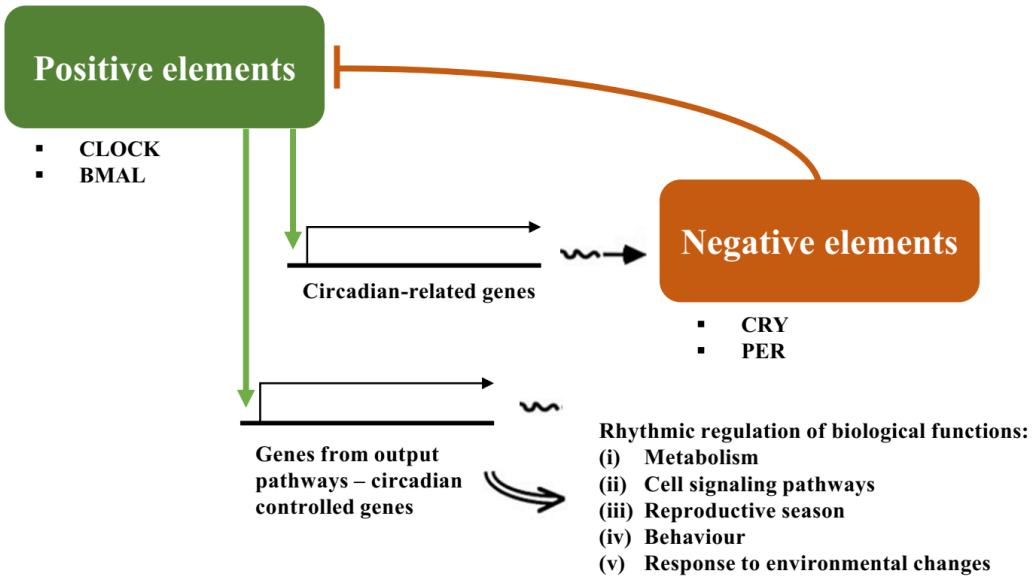

Figure 1.1. Overview of the core circadian system and output pathways [adapted from (Dunlap 1999)]

As previously stated, light is crucial for circadian system and organisms possess photoreceptor proteins (CRY and PER) that perceive the light-dark changes and properly generate input pathways to transduce the signal to the circadian core (Vallone et al. 2004; Tamai et al. 2007; Vatine et al. 2009). Two studies in zebrafish characterized a light-induced transcriptome and revealed several genes whose expression depends on light, revealing a multi-level regulation of circadian rhythms by light-cycles (Weger et al. 2011; Ben-Moshe et al. 2014). This light-dependence was also recently shown for the Atlantic cod *Gadus morhua* (Lazado et al. 2014). The molecular explanation remains in light-induced *cry1aa* and *per2* circadian genes that encode for negative core elements of the circadian transcription-translation loop with CRY1AA acting as a direct inhibitor of CLOCK:BMAL activation via preventing heterodimer formation (Ishikawa et al. 2002; Tamai et al. 2007) while PER2 plays a more complex role as either transcriptional coactivator or corepressor depending on its transcriptional regulatory targets (Wang et al. 2015). Photoreception is particularly interesting in fish as they exhibit several adaptations for this mechanism. Contrary to most vertebrates which only perceive light through the eyes, fish also possess a photosensitive pineal gland, dermal melanophores, and brain photoreceptors (Foulkes et al. 2016). In addition, fish possess independent peripheral photoreceptors and self-sustaining circadian oscillators, essentially in every tissue (Whitmore et al. 1998). Moreover, tissue and cell culture studies in zebrafish showed these peripheral clocks are directly entrained by light (Whitmore et al. 1998).

Circadian rhythms can also be entrained by temperature (Liu et al. 1998; Tsuchiya et al. 2003; Lahiri et al. 2005; Chappuis et al. 2013; Jerônimo et al. 2017). In mammals it was demonstrated that peripheral cells *in vitro* could sense the change of room temperature as a cue for entrainment of circadian system (Tsuchiya et al. 2003). In zebrafish, temperature has also an important role in circadian clock (Lahiri et al. 2005; Jerônimo et al. 2017), and it was proposed that temperature could entrain the phase of the system by driving expression levels of *per3*, and other clock genes (namely *cry2* and *cry1ba*) via an alternative hypothetical enhancer upstream the E-box (Lahiri et al. 2005). In this model, *per1b* (formerly known as *per4*) promoter integrates temperature and light regulatory input from the E- boxes together with regulation by other temperature-driven elements (Lahiri et al. 2005). Recently, it was found in a transcriptome profiling of two Iberian freshwater fish (*Squalius carolitertii* and *S. torgalensis*) two differentially expressed genes (*cry1aa* and *per1a*) between a control condition and a thermal stress condition (Jesus et al. 2016). Subsequent studies with the mentioned genes reinforced the dependence on temperature for proper entrainment of the circadian system in these species by showing a response in expression of *cry1aa* and *per1a* genes in *S. carolitertii* when subjected to a simulated scenario of moderate climate change (Jesus et al. 2017). Other studies connected the circadian system with stress response through a network involving heat-shock proteins (HSP). When exposed to thermal stress organisms increase up to 10% the expression of HSPs (Söti et al. 2005). This activation is related to heat-shock factors (HSFs) that bind regions in the DNA called heat-shock elements (HSE). In normal situations, HSP90 binds HSF1 and prevents its migration for the nucleus, but in thermal stress, these two proteins dissociate and HSF1 is free to migrate to the nucleus and binds to HSE sequence. It was demonstrated that this pathway can affect the expression of *per2* in mammals as this gene possesses in its promoter a heat-shock element (HSE) sequence (Kornmann et al. 2007; Tamaru et al. 2011; Chappuis et al. 2013). Other authors studying this same relation in zebrafish hypothesize that HSF1 activation is related to circadian regulation in at least two ways: (1) HSP activation improves the ability of the organism to deal with thermal stress, but HSP also interacts with BMAL1 to maintain the optimal function of the circadian system and (2) HSF1

activates other circadian-related genes as *per2* that may result in synchronization of the circadian machinery and maintain the outputs of the circadian system fully functional (Jerônimo et al. 2017). In the same work, Jerônimo et al. (2017) identified the channel TRPV1 (Transient receptor potential cation channel subfamily V member 1), a thermo-sensitive receptor, as an integrant part of zebrafish circadian rhythm by playing a role in heat-mediated *hsp90* and *per2* responses. By blocking the channel with a TRPV1 inhibitor, the authors demonstrated that even though *hsp90* expression was unaffected, *per2* expression was partially downregulated (Jerônimo et al. 2017).

Several studies have been conducted to understand the circadian system at its different levels of organization. In fish, these studies are particularly interesting once they are a very diverse group of animals adapted to nearly all aquatic environments and possess a larger number of circadian paralogs as compared to the other vertebrates (Toloza-Villalobos et al. 2015). Some of these studies cover the evolutionary relationships of the core-clock gene families and the mechanisms driving their molecular evolution (Wang 2008a, 2008b, 2009; Liu et al. 2015), but several key questions are still opened, namely the reason for the preservation of these higher number of paralogs when compared to other vertebrates (Toloza-Villalobos et al. 2015). Toloza-Villalobos et al. (2015) described the possible evolutionary history of circadian-related gene families and pointed to a possible event of subfunctionalization among *per* genes. Even though neofunctionalization was been a widely assumed fate for these duplicated circadian-related genes during evolution, there are still some opened questions concerning the mechanisms driving molecular evolution in these gene families, which is clearly noted in functional analysis as several of the paralogs retain ancestral functions creating a redundant pool of circadian genes.

The genus *Squalius* Bonaparte, 1837 is represented in Portuguese rivers by four known species (*S. carolitertii*, *S. pyrenaicus*, *S. torgalensis* and *S. aradensis*) distributed across a latitudinal cline (Figure 1.2). *S. carolitertii* (Doadrio, 1988) inhabits the northern rivers of Portugal, *S. pyrenaicus* (Günther 1868) occurs in the Central and Southern drainages (e.g. Tagus and Almargem), while sister species *S. torgalensis* and *S. aradensis* (Coelho et al. 1998) are confined to small basins in the Southwest of Portugal. This distribution gains special interest from an evolutionary perspective since Portugal is at the frontier between two contrasting climate types: the Atlantic in the Northern region that is characterized by mild temperatures, and the Mediterranean in the Southern region typified by high temperatures and droughts (Table S1). The photoperiod is also variable along this latitudinal cline, and typically daytime is longer in southern region of Portugal for about 10 to 15 minutes when compared to the northern region. Moreover, these basins present differences in water pH, as southern basins are typically more alkaline than northern ones (Table S1). These environmental differences associated to distribution of *Squalius* species turn them into an excellent model to study evolutionary adaptation to different environmental backgrounds, in particular for studying adaptation of the circadian system to different environmental conditions of light and temperature.

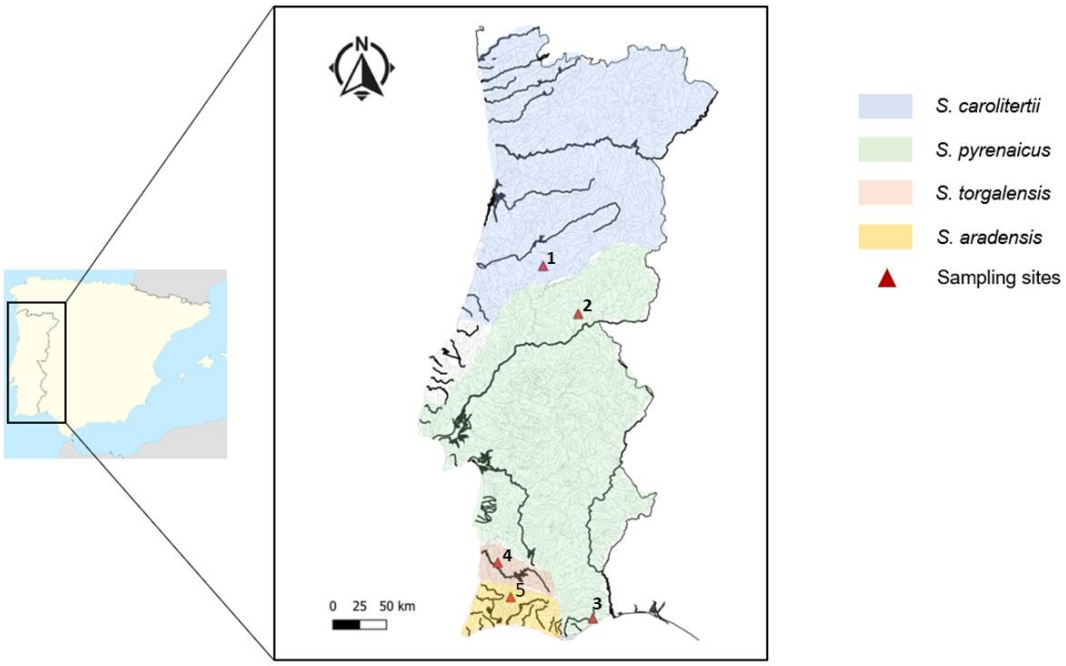

Figure 1.2. Spatial distribution of the four Portuguese *Squalius* species. Sampling sites are marked with red triangles: Mondego basin (1, Sótão river); Tagus basin (2, Ocreza river); Almargem basin (3, Almargem stream); Mira basin (4, Torgal stream); Arade basin (5, Odelouca stream).

Here we present an integrative study on molecular evolution of circadian system in Portuguese *Squalius* species. Transcriptomes of *S. carolitertii* and *S. torgalensis* previously published (Jesus et al. 2015) allowed the identification of circadian-related genes in these species by comparing them with transcriptomes of the reference organism *Danio rerio*. Phylogenetic analysis of gene families was performed using predicted protein sequences to elucidate evolutionary history of these families in the four species according the environmental gradient. Models to predict protein-protein interactions were used to investigate possible modifications in the function of duplicated genes and to provide insights on the fate of these duplications. Patterns of adaptation to the environmental conditions were studied using statistical tests to detect signatures of selection. However, as these tests do not have enough statistical power to detect all signatures of selection and do not identify impact of positively selected mutations in protein function, functional and structural features of the predicted proteins were inferred and compared between populations to better describe the mechanisms of adaptation of *Squalius* to the environmental cline of light and temperature. Specifically, physicochemical parameters as isoelectric point, instability and aliphatic index (see Table S6 for biological relevance), together with domains and post-translational modifications were predicted. Additionally, we characterised some features of TIMELESS in *Squalius* species, namely protein-protein interactions and family assignment using hidden-Markov models, with the purpose of providing insights on its function and clarify its relationship with circadian system.

2. Materials and Methods

2.1. Sampling

Muscle tissue from five wild adult fish of *S. carolitertii* and *S. torgalensis* species was already stored at -80ºC in RNAlater® (Ambion, Austin, TX, USA) from previous work (Jesus et al. 2017). They were sampled in Portuguese basins Mondego (40˚8’5.22"N; 8˚8’35.06"W) and Mira (37˚38'1.31"N; 8˚37’22.37"W), respectively. Samples of *S. pyrenaicus* were also stored at -80ºC from previous projects. This work includes two sampling sites: Almargem (37°09'50.7"N; 7°37'13.2"W) (Machado et al. 2016) and Tagus (39°43'48.2"N; 7°45'38.1"W) (Matos et al. 2016).

Squalius aradensis individuals were captured from Portuguese basin Arade (37º17'0.53"N; 8º29'7.31"W) under the license 421/2017/CAPT issued by Portuguese authority for Conservation of endangered species [ICNF (Instituto da Conservação da Natureza e das Florestas)]. After capturing, the specimens were transported alive to the laboratory in aerated containers and sacrificed upon arrival with an overdose of tricaine mesylate (400 ppm of MS-222; Sigma-Aldrich, St. Louis, MO, USA) with sodium bicarbonate (1:2) following the recommended ethical guidelines (ASAB/ABS, 2012) and European Union regulations. Efforts were made to minimize fish discomfort. Organs were stored in RNAlater® at -80ºC until further use. Distribution of the species and sampling sites are illustrated in Figure 1.2.

An environmental characterisation was done for each sampling site. Average water temperature and pH between 2001 and 2016 were retrieved from SNIRH [snirh.pt (National Information System of Water Resources); accessed in June 2018] and number of daylight hours was retrieved from Time and Date AS (timeanddate.com; accessed in June 2018). The data is summarised in table S1.

2.2. Identification of circadian system related genes in Iberian freshwater fish

Transcriptomes of *S. torgalensis* and *S. carolitertii,* already available and published (Jesus et al. 2015), were used to identify the circadian related genes in the study species. BLAST searches of both transcriptomes were conducted against two *Danio rerio* light-induced transcriptomes (Weger et al. 2011; Ben-Moshe et al. 2014) to identify potential genes related to the circadian system. An e-value threshold of 1×10^{-7} was used during the BLAST searches and only sequences with identity higher than 85% were retrieved to avoid the use of different duplicates or splicing isoforms. These light-responsive genes include several genes already characterised as components of the circadian system, but also several uncharacterised genes or genes unrelated to the circadian system. Due to this problem, and to avoid any false positive, an analysis of functional enrichment was accomplished. For this, a list of *Danio rerio* ENA accession numbers from the top blast results mentioned above was used to perform the enrichment analysis in DAVID functional annotation tool (Huang et al. 2007). An EASE score <0.05, a statistical test to examine the significance of gene–term enrichment with a modified Fisher's exact test, was used for all functional analyses performed in DAVID. Through this methodology we were able to find enriched GO terms among the genes retrieved. The most significant enriched GO terms for Biological Process and Molecular Function were filtered, first by rejecting all the genes with function unrelated to the circadian system, and among the circadian related genes using a threshold for adjusted p-values (Benjamini) of 0.05 to remove false positives. From the final list of genes only protein-coding genes whose function was related to the core circadian mechanism were maintained for further analysis (Table S2).

2.3. Gene sequencing and protein sequence prediction

Based on sequences retrieved from the transcriptomes, specific primers for Polymerase chain reactions (PCRs) were designed using PerlPrimer software v.1.1.19 (Marshall 2004) (Table S3) with the purpose of amplifying the same genes for all the studied species.

Total RNA was extracted from muscle samples of 25 individuals, 5 from each population. 1 mL TRI Reagent (Ambion, Austin, TX, USA) was added to 50–100 mg of muscle samples and, after homogenization with Tissue Ruptor (Qiagen, Valencia, CA, USA), RNA was extracted according to the TRI Reagent manufacturers protocol. TURBO DNase (Ambion, Austin, TX, USA) was employed to degrade any remaining genomic contaminants, followed by phenol/chloroform purification and LiCl precipitation (Cathala et al. 1983). Sample quality was checked using a NanoDrop™-1000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) based on the 260nm/280nm and 260nm/230nm absorbance ratios. Samples concentration were determined with Qubit® 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) to ensure enough quantity of homogeneous RNA for cDNA synthesis. Synthesis of cDNA was performed, according to manufacturer's protocol, using a RevertAid H Minus First Strand cDNA synthesis kit (Thermo Fisher Scientific, Waltham, MA, USA) and stored subsequently at -20˚C until further use. PCRs were performed in 25 μL reactions containing 10–100 ng of cDNA, 2 mM $MgCl_2$, 2 mM each dNTP, 10 μM each primer, Taq Polymerase (5 U/μL), and 1x Taq buffer using the thermocycler conditions described in Table S4. PCR products were confirmed using a 1% agarose gel electrophoresis, and after purification with ExoSAP-IT® PCR Product Cleanup (Affymetrix, Inc., Santa Clara, CA, USA), they were sequenced by Sanger sequencing.

Sequences were aligned and edited using Sequencher v.4.2 (Gene Codes Corp., Ann Arbor, MI, USA). Nucleotide sequences were deposited in European Nucleotide Archive (ENA) database under the accession numbers available in Tables 3.1, 3.3, 3.6, 3.9 and 3.12. CLC Sequence Viewer v.7.5. (CLC bio, Aarhus, Denmark) was used to predict protein sequences for *in silico* analysis. BLAST searches were conducted with resulting protein sequences against UniProt database (The UniProt Consortium 2015) to ensure their reliability. Protein sequences for *Danio rerio* were retrieved from UniProt database for each protein (Table S2), as well as *Drosophila melanogaster* homolog sequence for each gene family (Table S5). Protein sequences were aligned by gene family using the M-Coffee method, that combines several alignment algorithms (eg. MUSCLE, MAFFT and CLUSTAL) (Wallace et al. 2006) available in the T-Coffee web server (Di Tommaso et al. 2011). For each individual gene, nucleotide sequences of *Squalius* species were also aligned using M-Coffee.

2.4. Phylogenetic analysis

The most appropriate model for amino acid substitution for each data sets was determined with ProtTest v.3.0 (Abascal et al. 2005; Darriba et al. 2011), for both based on the Akaike information criteria and Bayesian information criteria. Phylogenetic trees were reconstructed for each gene family independently using Bayesian Inference method as implemented in MrBayes v.3.2.6 (Huelsenbeck & Ronquist 2001; Ronquist et al. 2012) using *D. melanogaster* protein sequences as outgroup. For *timeless* gene, a phylogenetic analysis was also done using the mentioned approach but using *Danio rerio* as outgroup. 500,000 generations of Monte Carlo Markov Chain (MCMC) were run using as

priors the parameters determined in ProtTest for protein sequences. Trees were sampled every 500 generations during the analysis. The first 50,000 generations were excluded as burn-in after examining the variation in log-likelihood scores over time. Phylogenetic trees were constructed using protein sequences instead of nucleotide sequences to avoid bias from the third codon rapid evolution. Protein sequences used correspond to the direct translation of nucleotide sequences under the standard genetic code. All trees were edited in FigTree v1.4.2 (A. Rambaut, University of Edinburgh, UK; http://tree.bio.ed.ac.uk/software/figtree/).

2.5. Analysis of signatures of selection

Signatures of selection were examined based on the dN/dS (known as ω) statistics using four models implemented in HyPhy (Kosakovsky Pond et al. 2005) through the Datamonkey adaptive evolution webserver (Kosakovsky Pond & Frost 2005; Weaver et al. 2018 ; http://www.datamonkey.org/; accessed in August 2018), including (1) BUSTED (Branch-site Unrestricted Statistical Test for Episodic Diversification) (Murrell et al. 2015) that provides a gene-wide test for positive selection; (2) MEME (Mixed Effects Model of Evolution) (Murrell et al. 2012) a mixed-effects maximum likelihood approach to test the hypothesis that individual sites have been subject to episodic positive selection; (3) aBSREL (adaptive Branch-Site Random Effects Likelihood) (Kosakovsky Pond et al. 2011; Smith et al. 2015) for genes whose a signals of positive selection were detected with BUSTED or MEME, to test if positive selection has occurred on a proportion of branches; and (4) FEL (Fixed Effects Likelihood) (Kosakovsky Pond & Frost 2005b) that uses a maximum likelihood approach to infer nonsynonymous (dN) and synonymous (dS) substitution rates on a per-site basis for a given coding alignment and corresponding phylogeny and tests the hypothesis that individual sites have been subject to pervasive positive or negative selection.

2.6. Functional analysis and structural organisation of the predicted proteins

Homology methods available on several resources at the ExPASy Server (Gasteiger et al. 2005) were used to infer several properties of the proteins. Specifically, physicochemical parameters of the proteins were predicted using ProtParam (Gasteiger et al. 2005). Parameters estimated were isoelectric point (pI), instability and aliphatic index, and its biological significance is described in Table S6. Differences in physicochemical parameters were tested statistically in R v.3.2.3 (R Core Team 2015). Data was checked for normality (Shapiro-Wilk's test). Due to lack of normality, a Kruskal-Wallis Rank Sum Test (KW) was performed to identify overall statistical differences in parameters across the populations. Pairwise Wilcoxon Rank Sum Tests were performed to compare the different groups and assess the origin of KW significance. Plots representing mean ± standard error for the five individuals per parameter and protein were constructed using the R package ggplot2 (Wickham 2016).

A sequence-based prediction of protein features (sequence domains and post-translational modification sites) was accomplished, using the online tool ScanProsite (de Castro et al. 2006 ; http://prosite.expasy.org/ accessed in March 2018) against the PROSITE database (Sigrist et al. 2002, 2013). Additional predictions of family assignment and sequence domains were accomplished using HMMER web server (Finn et al. 2011; Prakash et al. 2017; Potter et al. 2018 ; https://www.ebi.ac.uk/Tools/hmmer/) against Pfam (Finn et al. 2016), CATH-Gene3D (Dawson et al. 2017), TIGRFAMs (Haft et al. 2013), SUPERFAMILY (Oates et al. 2015) and PIRSF (Wu 2004) databases. These tools are mainly based on collections of Hidden-Markov Models to support the predictions (Eddy 1998, 2011). The number of predicted post-translational modification (PTM) sites

were compared between species for each protein to infer about possible modifications in protein regulatory mechanisms. Representative images of structural organization of domains and locations of sites under selection were created and edited in PROSITE tool MyDomains (Hulo et al. 2008).

Additionally, a sequence-based prediction of protein-protein interactions (PPI) was done using the online tool STRING (von Mering et al. 2003; Szklarczyk et al. 2015 ; https://string-db.org/cgi/input.pl accessed in August 2018). Interactors were predicted comparing *Squalius* predicted protein sequences against fish PPI databases using a threshold for score of 0.7.

3. Results

Analysis of *S. torgalensis* and *S. carolitertii* transcriptomes (Jesus et al. 2015) and comparison with light-induced zebrafish transcriptomes revealed four main gene families involved (Cryptochromes, Period, CLOCK and BMAL) in the circadian system, and one single-copy gene (*timeless*) whose main function was related to the circadian system. These gene families correspond to those already described for other fish species, specifically *Danio rerio* (Wang 2008a, 2008b, 2009; Liu et al. 2015). These identified genes were re-sequenced in *S. torgalensis* and *S. carolitertii* and sequenced *de novo* for *S. pyrenaicus* and *S. aradensis.* Non-redundant sequences were deposited in ENA under the accession numbers presented in Tables 3.1, 3.3, 3.6, 3.9 and 3.12. For CLOCK family we were unable to identify *clock2* gene, previously characterized in other fish species, but identified two protein-coding genes (*clocka* and *clockb*). The most probable reason for this may be related to the lack of information in the transcriptome due to sequencing limitations or low expression levels of this gene in sequenced organs.

3.1. Evolution and characterisation of Cryptochrome Family

Six *cry* genes encoding for six proteins were identified (Table 3.1) and classified in three major groups (CRY1, CRY2 and CRY3) according to phylogenetic analysis (Figure 3.1). Among CRY1 group, four protein-coding genes were identified: *cry1aa, cry1ab, cry1ba* and *cry1bb.* CRY2 and CRY3 groups possess only one protein-coding gene each: *cry2* and *cry3,* respectively. Phylogenetic analysis revealed a possible evolutionary convergence between *S. aradensis* and Almargem population of *S. pyrenaicus* pointing to possible functional differences between the two populations of *S. pyrenaicus* (Tagus and Almargem).

Table 3.1. *Cryptochrome* genes identified with respective annotations obtained in functional annotation analysis. ENA accession numbers are for *Squalius* sequences obtained by Sanger sequencing in this work

		Gene ontology annotation		
Gene symbol	**Gene name**	**Biological process**	**Molecular function**	**ENA accession numbers**
cry1aa	cryptochrome circadian clock 1aa	· entrainment of circadian clock · response to light stimulus · negative regulation of transcription · response to hydrogen peroxide	· protein binding	LS999803 to LS999807
cry1ab	cryptochrome circadian clock 1ab	· negative regulation of transcription	· non-annotated	LS999808 to LS999812
cry1ba	cryptochrome circadian clock 1ba	· response to light stimulus · response to temperature stimulus · negative regulation of transcription	· non-annotated	LS999813 to LS999817
cry1bb	cryptochrome circadian clock 1bb	· negative regulation of transcription	· non-annotated	LS999818 to LS999822
cry2	cryptochrome circadian clock 2	· response to light stimulus · response to temperature stimulus · negative regulation of transcription	· non-annotated	LS999666 to LS999670
cry3	cryptochrome circadian clock 3	· signal transduction	· photoreceptor activity	LS999671 to LS999675

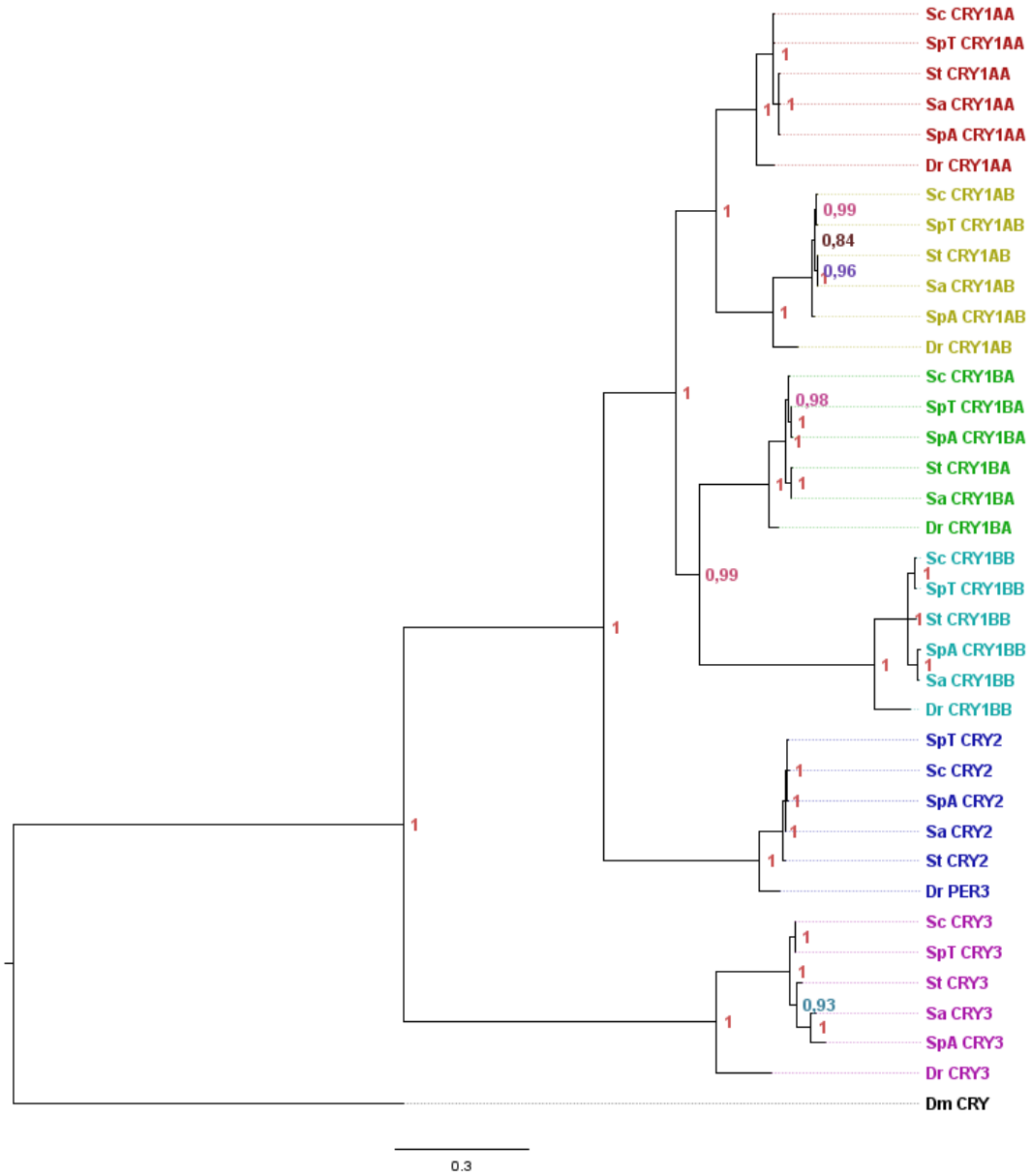

Figure 3.1. A phylogenetic tree constructed by the Bayesian Inference method for CRY proteins with fly CRY as outgroup using the LG substitution model (Le & Gascuel 2008) with a discrete Gamma distribution (+G) with 5 rate categories. Values on branch nodes represent Bayesian posterior probabilities. Sc, *Squalius carolitertii;* SpT, *Squalius pyrenaicus* (Tagus population)*;* St, *Squalius torgalensis;* Sa, *Squalius aradensis;* SpA, *Squalius pyrenaicus* (Almargem population)*;* Dr, *Danio rerio;* Dm, *Drosophila melanogaster*.

For *cry* genes no signals of gene-wide positive selection were found. Even though some indications of gene-wide positive selection in *cry1aa* gene were found, these were not statistically significant. However, one site subjected to episodic positive selection in *cry1aa* gene was identified (Table S7). This site corresponds to a non-conservative non-synonymous mutation from a lysine to a glutamine (K275Q) in *S. torgalensis* and may have an impact in the overall charge of the protein due to a change from an uncharged to a charged amino acid. For all *cry* genes, several sites were found to be under pervasive negative selection (Table S8). Important to denote that several of these sites under negative selection are located inside functional domains of CRY proteins (Figure 3.2).

Table 3.2. Analysis of gene-wide positive selection in *cry* genes using the BUSTED method implemented in Datamonkey webserver. A threshold of 0.1 was used for statistical significance.

Gene	Model	Parameters	Likelihood (lnL)	AICc	LRT	P-value	ω1	ω2	ω3
cry1aa	Unconstrained model	22	-2696.9	5438.4	-3,4	0.191	0.00 (93.01%)	0.17 (5.65%)	10.61 (1.35%)
	Constrained model	21	-2698.6	5439.6			0.00 (7.14%)	0.00 (84.77%)	1.00 (8.08%)
cry1ab	Unconstrained model	22	-2563.6	5171.8	0	1,000	0.37 (39.59%)	1.00 (60.41%)	1.58 (0.00%)
	Constrained model	21	-2563.6	5169.8			0.38 (39.08%)	0.99 (60.92%)	1.00 (0.00%)
cry1ba	Unconstrained model	26	-2834.9	5722.2	-0,4	0.772	1.00 (44.69%)	1.00 (35.53%)	4.01 (19.77%)
	Constrained model	25	-2835.1	5720.7			1.00 (19.82%)	1.00 (44.65%)	1.00 (35.53%)
cry1bb	Unconstrained model	26	-2978.5	6009.5	0	1,000	0.33 (35.81%)	0.45 (64.19%)	1.03 (0.00%)
	Constrained model	25	-2978.5	6009.5			0.33 (35.81%)	0.45 (64.19%)	1.03 (0.00%)
cry2	Unconstrained model	24	-2716.4	5481.3	0	1,000	0.06 (55.67%)	0.09 (44.33%)	1.00 (0.00%)
	Constrained model	23	-2716.4	5479.2			0.06 (55.67%)	0.09 (44.33%)	1.00 (0.00%)
cry3	Unconstrained model	24	-2670.3	5389.2	0	1,000	0.37 (0.00%)	0.44 (100.00%)	1.00 (0.00%)
	Constrained model	23	-2670.3	5386.7			0.37 (0.00%)	0.44 (100.00%)	1.00 (0.00%)

Concerning physicochemical parameters, all CRY proteins presented significant changes (Figure S1 – S6). Changes in protein pI were detected in CRY1AA (KW; $p<0.001$) between *S. torgalensis* and other species ($p<0.001$). Changes in pI were also detected in CRY1AB where southern species *S. torgalensis* and *S. aradensis* have lower pI when compared to all the other species ($p<0.01$), in CRY1BA where pI was higher for Almargem population of *S. pyrenaicus* in relation to other populations ($p<0.01$), and in CRY1BB where both populations of *S. pyrenaicus* tend to have lower pI values ($p<0.05$). CRY2 and CRY3 proteins also present changes in pI (KW; $p<0.01$), specifically in CRY2 pI for which *S. aradensis* has a lower pI than the other species ($p<0.01$), and in CRY3 pI for which *S. aradensis* and Almargem population of *S. pyrenaicus* show a lower value for pI ($p<0.01$ and $p<0.001$, respectively), while *S. torgalensis* presents the higher value for this parameter ($p<0.01$). Concerning instability, CRY1AA from *S. torgalensis* presents an increased instability compared to other species ($p<0.01$), as well as *S. aradensis* and Almargem population of *S. pyrenaicus* when compared to *S. carolitertii* and *S. pyrenaicus* from Tagus ($p<0.05$). For CRY1AB Almargem population of *S. pyrenaicus* presents an increased instability compared to other populations ($p<0.01$). CRY1BA from *S. aradensis* and *S. torgalensis* presents patterns of higher instability when compared with CRY1BA from other species ($p<0.001$), while both populations of *S. pyrenaicus* display patters of higher stability compared to the other species ($p<0.001$). CRY1BB tends to have a higher instability in both populations of *S. pyrenaicus* ($p<0.001$). For CRY2, both *S. torgalensis* and *S. aradensis* have signals of higher instability ($p<0.01$), with *S. torgalensis* having the higher value of instability, and, for CRY3 southern populations of *S. aradensis* and *S. pyrenaicus* present the lower values ($p<0.01$) while northern populations of *S. carolitertii* and *S. pyrenaicus* present the higher values ($p<0.01$). *S. torgalensis* presents, for CRY3 instability, and intermediate value when compared to the other species ($p<0.01$). Aliphatic index is modified in all CRY proteins (KW; $p<0.01$). For CRY1AA which south populations displayed lower values when compared to northern populations ($p<0.05$), for CRY1AB *S. carolitertii* and Tagus population of *S. pyrenaicus* present a slightly higher aliphatic index when compared to other populations ($p<0.05$), for CRY1BA both populations of *S. pyrenaicus* display patterns of higher aliphatic index ($p<0.01$) compared to the other species, and, for CRY1BB *S. torgalensis* displays a lower value ($p<0.01$), and the same trend can be observed for the Almargem population of *S. pyrenaicus* ($p<0.01$). A lower value of aliphatic index was detected for CRY2 from *S. aradensis* ($p<0.05$), and for CRY3 aliphatic index, *S. pyrenaicus* form Almargem has the higher value

(p<0.01) while both *S. carolitertii* and *S. pyrenaicus* from Tagus have the lower values (p<0.01). *S. aradensis* and *S. torgalensis* have intermediate values for CRY3 aliphatic index (p<0.01).

When comparing PTM for CRY protein several modifications were detected with possible impact in protein function and its regulation (Table S9). Phosphorylation is the PTM with more changes in these proteins. For CRY1AB we found an extra protein kinase C phosphorylation site in *S. pyrenaicus* from Almargem, and an extra casein kinase II phosphorylation site in *S. torgalensis* and *S. aradensis.* For CRY1BA, an extra cAMP- and cGMP-dependent protein kinase phosphorylation site was found in *S. aradensis* and *S. torgalensis*, while in CRY3 protein *S. torgalensis* and *S. aradensis* lack this phosphorylation site and all the other species present one possible cAMP- and cGMP-dependent protein kinase phosphorylation site. CRY1BB and CRY3 have modification in protein kinase C phosphorylation sites in all the species (Table S9). Concerning other PTMs, CRY1BB has a modification with *S. torgalensis* and *S. pyrenaicus* from Almargem lacking a N-glycosylation site. Additionally, the lack for ATP/GTP binding site in CRY1BA of *S. carolitertii* and both populations of *S. pyrenaicus*, that may be related to the lack of one cAMP- and cGMP-dependent protein kinase phosphorylation site in this protein for the mentioned species.

A structural analysis of CRY proteins revealed the presence of two common domains to all CRY proteins (Figure 3.2): the Photolyase/cryptochrome alpha/beta domain (PHR) and the FAD-binding domain (FAD_binding). For most CRY proteins, sites detected to be under pervasive negative selection are in domain regions, with a few exceptions whose sites are in other regions of the proteins.

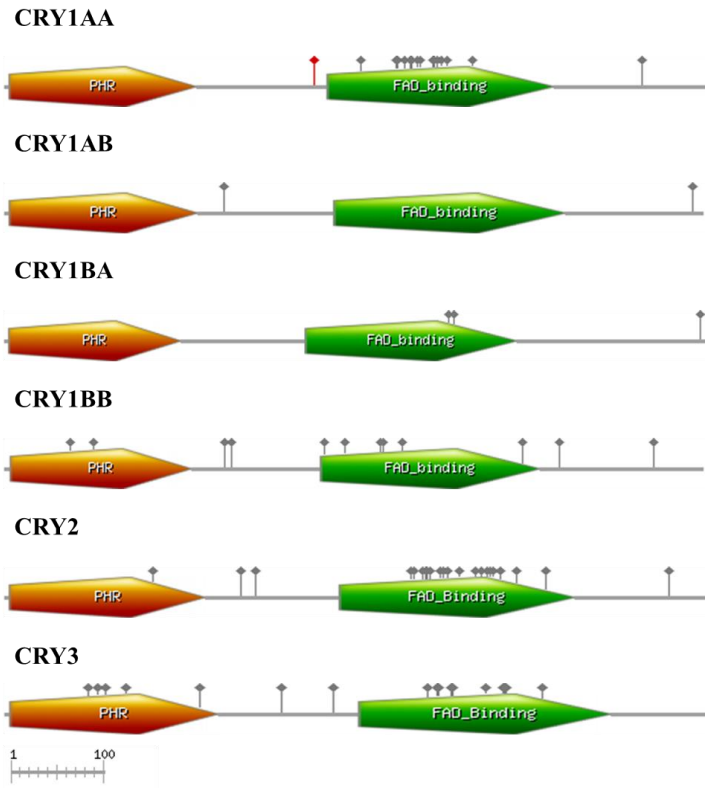

Figure 3.2. Schematic diagrams of the full length *Squalius* CRY1AA, CRY1AB, CRY1BA, CRY1BB, CRY2, and CRY3 proteins. In orange is represented the photolyase-like domain (PHR) and in green is represented the FAD-binding domain. Red dots represent sites under episodic positive selection and dots in grey represent sites under negative selection.

Analysis of protein-protein interactions revealed interaction of CRY proteins mainly with other circadian-related protein here studied (Table S10). However, some other PPIs were found and considered to be relevant in the context of this work. Both CRY1AA, CRY1AB, CRY1BA and CRY2 were found to interact with bHLHe41 (basic helix-loop-helix family, member e41), a protein annotated with GO terms "sleep" and "response to cold", showing a relation with both the circadian rhythm and response to thermal conditions. CRY1AA was found to interact with TEFa (thyrotrophic embryonic factor alpha) annotated with the term "cellular response to light stimulus". CRY1BB was found to interact with NR1D2b (nuclear receptor subfamily 1, group D, member 2b) annotated with the term "regulation of circadian rhythm". CRY2 was found to interact with two proteins related to the circadian rhythm: NR1D1 (nuclear receptor subfamily 1, group D, member 1) and CIARTA (circadian-associated repressor of transcription a) both annotated with the term "circadian regulation of gene expression". Additionally, CRY2 was found to interact with HSF2 (heat-shock transcription factor 2), a transcription factor responsible for activation of expression of HSP proteins in response to thermal stress. CRY3 was found to interact also with NFIL3-5 (nuclear factor, interleukin 3-regulates, member 5) annotated with the term "circadian regulation of gene expression".

3.2. Evolution and characterisation of Period Family

For the PER family, four protein-coding genes were identified (Table 3.3), and encoded proteins belong to three different clades according to phylogenetic analysis (Figure 3.3). PER1 clade comprises two proteins (PER1A and PER1B), and consequently the corresponding genes (*per1a* and *per1b*), while PER2 and PER3 clades are composed each by a single protein, encoded by the genes *per2* and *per3*, respectively. All these genes were previously identified in *D. rerio* and other fish species (Wang 2008a).

Table 3.3. *Period* genes identified with respective annotations obtained in functional annotation analysis. ENA accession numbers are for *Squalius* sequences obtained by Sanger sequencing in this work

		Gene ontology annotation		
Gene symbol	**Gene name**	**Biological process**	**Molecular function**	**ENA accession numbers**
per1a	period circadian clock 1a	· photoperiodism	· transcription corepressor binding	LS999650 to LS999654
per1b	period circadian clock 1b	· regulation of circadian system · response to light stimulus · response to temperature system · visual perception	· chromatin binding · transcription corepressor binding · transcription regulatory region sequence-specific DNA binding	LS999655 to LS999659
per2	period circadian clock 2	· entrainment of circadian clock by photoperiod · response to light stimulus · response to cold · response to UV radiation · positive regulation of circadian rhythm	· chromatin binding · transcription corepressor binding · transcription regulatory region sequence-specific DNA binding	LS999660 to LS999665
per3	period circadian clock 3	· circadian regulation of gene expression · photoperiodism	· transcription corepressor binding	LS999676 to LS999681

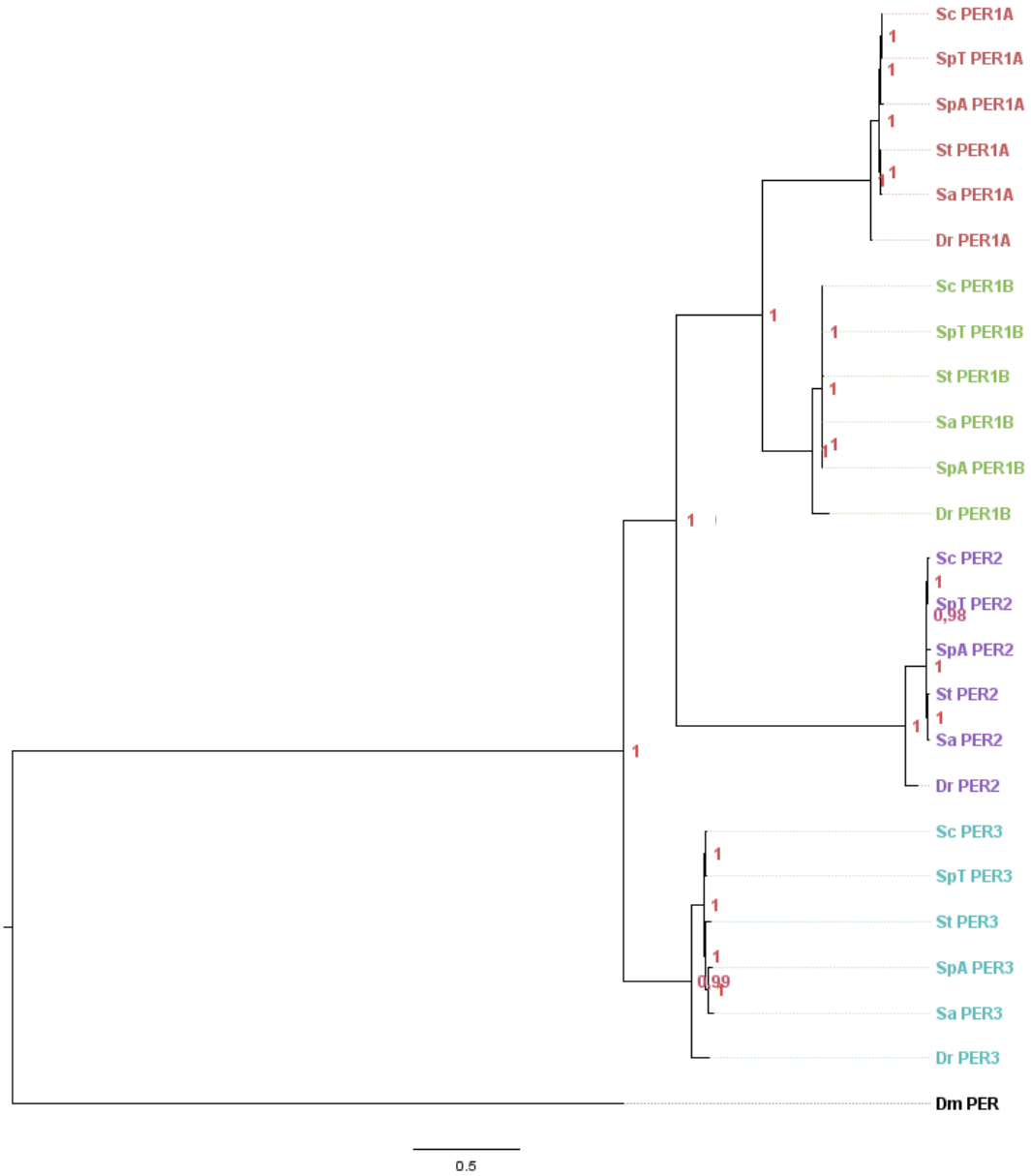

Figure 3.3. A phylogenetic tree constructed by the Bayesian Inference method for PER proteins with fly PER as outgroup using the JTT substitution model (Jones et al. 1992) with empirical amino acid frequencies from the data (+F). Values on branch nodes represent Bayesian posterior probabilities. Sc, *Squalius carolitertii;* SpT, *Squalius pyrenaicus* (Tagus population); St, *Squalius torgalensis;* Sa, *Squalius aradensis;* SpA, *Squalius pyrenaicus* (Almargem population); Dr, *Danio rerio;* Dm, *Drosophila melanogaster*.

Among PER genes, both *per2* and *per3* displayed signals of gene-wide positive selection (Table 3.4). For *per2*, *S. torgalensis* and *S. aradensis* were both found to be under positive selection, while for *per3* signature of positive selection was found in Almargem population of *S. pyrenaicus* (Table 3.X). Additionally, for *per1b* signals of positive selection in *S. torgalensis* was found in a branch analysis, even though no signals of gene-wide positive selection were found for this gene. In site analysis, sites under episodic positive selection were found in all *per* genes (Table S7). For PER2, mutations T402D and G1208T, and for PER3, mutation K429Y, were found to be located inside functional domains. All these sites correspond to non-conservative mutations. All *per* genes present a considerable number of positions under pervasive negative selection, specifically *per2* with 59 sites identified (Table S11).

Table 3.4. Summary of gene-wide positive selection analysis in *per* genes using the BUSTED method implemented in Datamonkey webserver. A threshold of 0.1 was used for statistical significance. Rows shaded in grey correspond to genes whose test for positive selection was statically significant.

Gene	Model	Parameters	Likelihood (lnL)	AICc	LRT	P-value	ω1	ω2	ω3
per1a	Unconstrained model	22	-2696.9	5438.4	-3,4	0.191	0.00 (71.49%)	0.18 (19.04%)	15.18 (9.47%)
	Constrained model	21	-2698.6	5439.6			0.01 (0.00%)	1.00 (84.70%)	1.00 (15.30%)
per1b	Unconstrained model	22	-2563.6	5171.8	0	1,000	0.09 (10.11%)	0.12 (89.85%)	317.33 (0.04%)
	Constrained model	21	-2563.6	5169.8			0.00 (86.12%)	0.16 (0.00%)	1.00 (13.88%)
per2	Unconstrained model	26	-7327.3	14706.8	-21	0,000	0.00 (2.26%)	0.32 (96.80%)	22.65 (0.94%)
	Constrained model	25	-7337.8	14725.8			0.00 (59.86%)	0.36 (0.00%)	1.00 (40.14%)
per3	Unconstrained model	26	-5733.9	11520,0	-7,4	0.024	0.00 (81.58%)	0.00 (13.60%)	9.88 (4.82%)
	Constrained model	25	-5737.6	11525.4			0.00 (61.26%)	0.23 (0.00%)	1.00 (38.74%)

Table 3.5. Summary of branch-site positive selection analysis in *per* genes using the aBSREL method implemented in Datamonkey webserver. A threshold of 0.1 was used for statistical significance. Rows shaded in grey correspond results whose test for positive selection was statically significant.

Protein	Branch	Optimized branch length	LRT	P-value	ω distribution over sites
PER1A	*S.carolitertii*	0.0003	0.530	0.324	ω1 = 10000000000 (100%)
	S.pyrenaicus (Tagus)	0.0000	0.000	1.000	ω1 = 1.00 (97%)
	S.torgalensis	0.0008	0.000	1.000	ω1 = 0.393 (100%)
	S.aradensis	0.0011	2.467	0.111	ω1 = 10000000000 (100%)
	S.pyrenaicus (Almargem)	0.0037	0.025	0.471	ω1 = 1.11 (100%)
PER1B	*S. carolitertii / S. pyrenaicus (Tagus)*	0.0019	0.000	1.000	ω1 = 0.00 (100%)
	S. torgalensis	0.0017	8.525	0.025	ω1 = 0.00 (99.7%); ω2 =226 (0.34%)
	S. aradensis	0.0002	0.000	1.000	ω1 = 0.00 (100%)
	S. pyrenaicus (Almargem)	0.0000	0.000	1.000	ω1 = 1.00 (97%)
PER2	*S.carolitertii*	0.0005	0.000	1.000	ω1 = 0.00 (100%)
	S.pyrenaicus (Tagus)	0.0017	0.118	1.000	ω1 = 1.48 (100%)
	S.torgalensis	0.0176	18.760	0.000	ω1 = 0.162 (97%); ω2 = 19.7 (2.7%)
	S.aradensis	0.0163	10.962	0.009	ω1 = 0.239 (100%); ω2 = 10000 (0.28%)
	S.pyrenaicus (Almargem)	0.0089	0.000	1.000	ω1 = 0.918 (100%)
PER3	*S.carolitertii*	0.0026	0.000	1.000	ω1 = 0.413 (100%)
	S.pyrenaicus (Tagus)	0.0006	0.000	1.000	ω1 = 1.00 (100%)
	S.torgalensis	0.0003	0.000	1.000	ω1 = 1.00 (100%)
	S.aradensis	0.0016	0.000	1.000	ω1 = 0.396 (100%)
	S.pyrenaicus (Almargem)	0.0062	8.105	0.043	ω1 = 0.0000000000000108 (97%); ω2 = 32.2 (3.2%)

All PER protein presented modifications on physicochemical parameters (Figure S7 – S10). For pI, changes were detected in PER1A (KW; $p<0.01$), between *S. pyrenaicus* from Almargem and the other populations ($p<0.05$) and in PER1B some marginal differences were also found for this parameter (KW; $p=0.056$). For PER2, changes were also detected in pI (KW; $p<0.001$), specifically with southern species *S. torgalensis* and *S. aradensis* presenting the lower values ($p<0.01$), with *S. torgalensis* having a slightly higher value than *S. aradensis* ($p<0.05$). Higher values of PER3 pI were detected for *S. torgalensis, S. aradensis* and *S. pyrenaicus* from Almargem ($p<0.01$). For instability, changes were detected in PER1A (KW; $p<0.01$) where *S. torgalensis* presents an increased value ($p<0.01$) and *S. pyrenaicus* from Almargem presents the lower value for this property ($p<0.01$). Instability remains, for the other species' PER1A approximately constant with an intermediate value between *S. torgalensis* and *S. pyrenaicus* from Almargem. PER1B presents small differences in instability. PER2 and PER3 both present significant changes in instability (KW; $p<0.01$). PER2

presents higher instability in *S. aradensis*, *S. torgalensis* and *S. pyrenaicus* from Almargem when compared to the northern populations of *S. carolitertii* and *S. pyrenaicus* ($p<0.01$), and in PER3 instability was found to be higher in *S. carolitertii* ($p<0.01$), being also slightly higher for both populations of *S. pyrenaicus* compared to the populations of *S. torgalensis* and *S. aradensis* ($p<0.05$). For aliphatic index, differences were found in all PER proteins (KW; $p<0.01$). In PER1A, *S. carolitertii* and *S. pyrenaicus* from Tagus display a considerable higher value of aliphatic index ($p<0.01$) while *S. aradensis* presents the lower value when compared to *S. torgalensis* and *S. pyrenaicus* from Almargem ($p<0.05$). In PER1B, *S. aradensis* and *S. pyrenaicus* from Almargem present a relatively lower value compared to other populations ($p<0.01$), and a slightly higher value was detected for *S. torgalensis* when compared to *S. carolitertii* and the Tagus population of *S. pyrenaicus* ($p<0.05$). For PER2, higher aliphatic index was found in *S. aradensis*, *S. torgalensis* and *S. pyrenaicus* from Almargem when compared to the northern populations of *S. carolitertii* and *S. pyrenaicus* ($p<0.01$). Important to denote the highest value for PER2 aliphatic index in *S. aradensis* ($p<0.001$). For PER3, aliphatic index was found to be higher in southern populations, more specifically in *S. torgalensis* ($p<0.001$) which presented the highest value; *S. aradensis* and *S. pyrenaicus* from Almargem also presented a relatively higher aliphatic index compared to northern populations ($p<0.01$). The lowest value for PER3 aliphatic index was found to be in *S. carolitertii* ($p<0.01$).

Comparing PTM sites for PER proteins several modifications were detected between species (Table S12). For PER1A modifications were detected in phosphorylation sites. While *S. aradensis* and *S. pyrenaicus* from Almargem presented the lack on one protein kinase C phosphorylation site when compared to the other populations, *S. pyrenaicus* from Almargem presented an extra casein kinase II phosphorylation site. Additionally, one extra cAMP- and cGMP-dependent protein kinase phosphorylation site was found in PER1A from *S. torgalensis* and *S. aradensis*. PER1A also presented modifications in N-glycosylation sites. For PER1B, the lack of one casein kinase II phosphorylation site in *S. aradensis* was also detected. PER2 and PER3 presented several alterations in PTM sites. In PER2 protein there were detected several alterations in phosphorylation sites, either protein kinase C and casein kinase II phosphorylation sites. To denote the lack of one tyrosine kinase phosphorylation site in *S. aradensis* and *S. pyrenaicus* from Almargem, and the lack of one cAMP- and cGMP-dependent protein kinase phosphorylation site in *S. torgalensis*. In PER3, modifications were found in phosphorylation sites, namely for protein kinase C phosphorylation sites, for which southern populations present extra sites when compared to the northern populations. PER3 from *S. pyrenaicus* from Almargem lacked one of the casein kinase II phosphorylation sites. Furthermore, PER3 from both populations of *S. pyrenaicus* lacks a N-glycosylation site compared to the other species.

A structural analysis revealed the presence of two domains common to all PER protein: Period-Arnt-Sim (PAS_3/PAS_11) domain and the Period protein 2/3C-terminal region domain (Figure 3.4). PER2 presents a PAS domain relatively conserved, but we found one of the sites under episodic positive selection to be located inside the PAS domains. This site corresponds to a mutation in *S. aradensis* and *S. pyrenaicus* from Almargem from a threonine to an aspartate (T402D) that could change the properties of the domain by changing from an uncharged amino acid to an acid charged amino acid. On the other site, most sites were found to be under pervasive negative selection in PER2 are located inside the PAS domain (Figure 3.4).

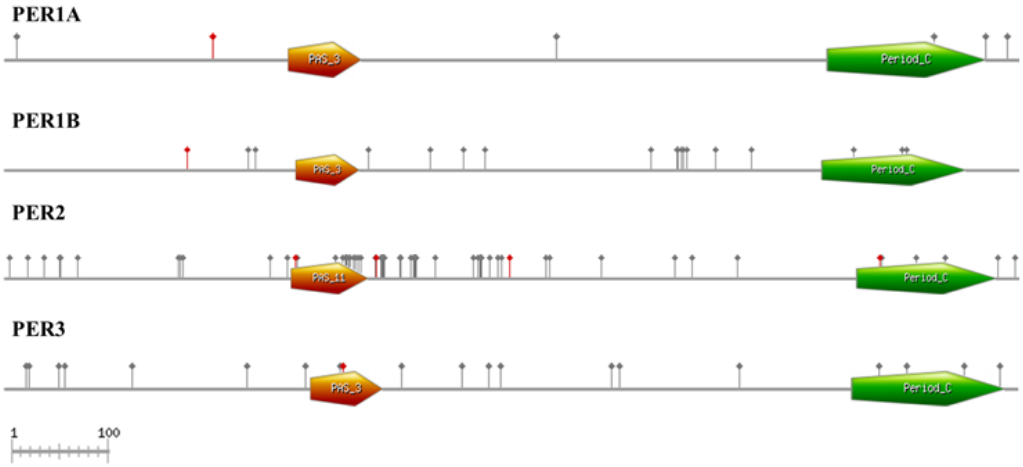

Figure 3.4. Schematic diagrams of the full length *Squalius* PER1A, PER1B, PER2 and PER3 proteins. In orange is represented the Period-Arnt-Sim (PAS_3/11) domain and in green the Period protein 2/3C-terminal region domain. Red dots represent sites under episodic positive selection and dots in grey represent sites under negative selection.

Protein-protein interactions in PER family were found for PER1A, PER1B and PER2 with bHLHe41 as we found in CRY proteins (Table S13). For PER1A and PER1B we also found interactions with CSNK1Db (casein kinase I, isoform delta b) annotated with the term "regulation of circadian rhythm. We also found interaction with other protein, CRY5 (cryptochrome 5), not directly related to the circadian system, but related to the Cryptochrome family and with DNA (6-4) photolyase activity, responsible for repair DNA damages induced by UV light.

3.3. Evolution and characterisation of CLOCK family

For CLOCK family two genes encoding for two protein belonging to a single clade according to phylogenetic analysis (Figure 3.5). CLOCK1 clade encompasses the two protein-coding genes (*clocka* and *clockb*) previously identified in *D. rerio* and other fish species (Wang, 2009). As previously mentioned, we were unable to identify *clock2* gene (already characterized in *D. rerio).*

Table 3.6. *Clock* genes identified with respective annotations obtained in functional annotation analysis. ENA accession numbers are for *Squalius* sequences obtained by Sanger sequencing in this work

		Gene ontology annotation		
Gene symbol	**Gene name**	**Biological process**	**Molecular function**	**ENA accession numbers**
clocka	clock circadian clock regulator a	· circadian regulation of gene expression · circadian rhythm · photoperiodism · positive regulation of gene expression · positive regulation of transcription · response to temperature stimulus · response to light stimulus	· DNA binding · protein binding · protein dimerization activity · DNA binding transcription factor activity, RNA polymerase II-specific	LS999682 to LS999686
clockb	clock circadian clock regulator b	· circadian rhythm · photoperiodism · regulation of transcription · response to light stimulus	· DNA binding · protein binding · protein dimerization activity	LS999687 to LS999691

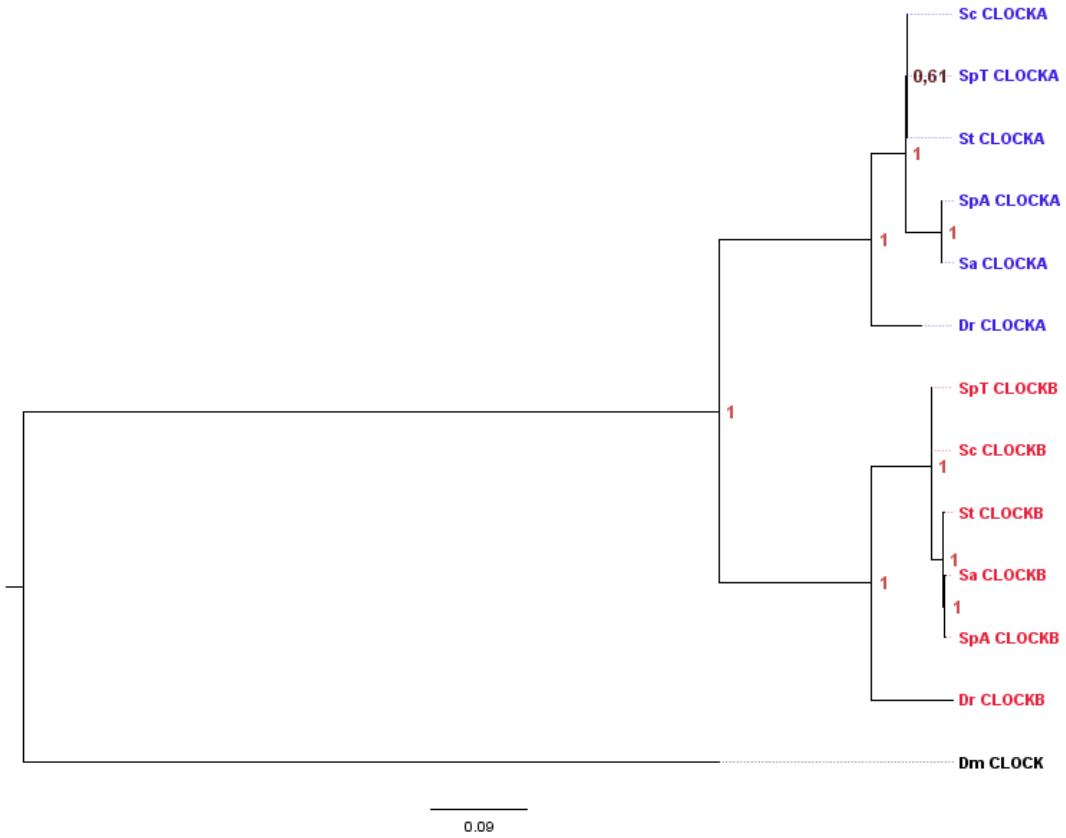

Figure 3.5. A phylogenetic tree constructed by the Bayesian Inference method for CLOCK proteins with fly CLOCK as outgroup using JTT substitution model (Jones et al. 1992) with empirical amino acid frequency (+F) using a discrete Gamma distribution (+G) with 5 rate categories. Values on branch nodes represent Bayesian posterior probabilities. Sc, *Squalius carolitertii*; SpT, *Squalius pyrenaicus* (Tagus population); St, *Squalius torgalensis*; Sa, *Squalius aradensis*; SpA, *Squalius pyrenaicus* (Almargem population); Dr, *Danio rerio*; Dm, *Drosophila melanogaster*.

Among CLOCK genes, both *clocka* and *clockb* displayed signals of gene-wide positive selection (Table 3.7). A branch analysis revealed signals of positive selection in *S. aradensis* and *S. pyrenaicus* (Almargem) for *clocka*, while for *clockb* the branch analysis revealed signals of positive selection in *S. carolitertii* (Table 3.8). No sites under episodic or pervasive positive selection were detected in *clock* genes (Table S7). Fifteen sites in *clocka* and five in *clockb* were found to be under negative selection (Table S14).

Table 3.7. Summary of gene-wide positive selection analysis in *clock* genes using the BUSTED method implemented in Datamonkey webserver. A threshold of 0.1 was used for statistical significance. Rows shaded in grey correspond to genes whose test for positive selection was statically significant.

Gene	Model	Parameters	Likelihood (lnL)	AICc	LRT	P-value	ω1	ω2	ω3
clocka	Unconstrained model	22	-3898.8	7842.1	-100,2	≤0.05	0.02 (89.13%)	0.05 (7.83%)	304.13 (3.03%)
	Constrained model	21	-3948.9	7940.2			0.00 (25.88%)	0.00 (53.43%)	1.00 (20.69%)
clockb	Unconstrained model	22	-3463.8	6912,0	-36,4	≤0.05	0.00 (4.11%)	0.00 (95.25%)	276.06 (0.65%)
	Constrained model	21	-3482	7006.4			0.00 (78.24%)	0.23 (0.00%)	1.00 (21.76%)

Table 3.8. Summary of branch-site positive selection analysis in *clock* genes using the aBSREL method implemented in Datamonkey webserver. A threshold of 0.1 was used for statistical significance. Rows shaded in grey correspond results whose test for positive selection was statically significant.

Protein	Branch	Optimized branch length	LRT	P-value	ω distribution over sites
CLOCKA	*S.carolitertii / S.pyrenaicus (Tagus)*	0.0004	0.000	1.000	ω1=0 (100%)
	S.torgalensis	0.0000	0.000	1.000	ω1=1 (100%)
	S.aradensis / S.pyrenaicus (Almargem)	0.0290	106.172	0.000	ω1=0.0310 (97.8%); ω2=451 (3.2%)
CLOCKB	*S.carolitertii*	0.0066	38.108	0.000	ω1 = 0.407 (99%); ω2 = 458 (0.62%)
	S.pyrenaicus (Tagus)	0.0004	0.598	0.620	ω1 = 10000000000 (100%)
	S.torgalensis / S.aradensis / S.pyrenaicus (Almargem)	0.0037	0.000	1.000	ω1 = 0.0462 (100%)

Considering physicochemical parameters, pI presented significant changes only in CLOCKB (Figure S12) protein (KW; $p<0.001$), specifically in *S. torgalensis* protein which has the higher value ($p<0.01$), *S. aradensis* and *S. pyrenaicus* from Almargem both have intermediate values when compared to the northern populations ($p<0.01$). For instability significant changes were detected for both CLOCKA (KW; $p<0.001$) and CLOCKB (KW; $p<0.001$). For CLOCKA (Figure S11), higher instability was found in *S. aradensis* and *S. pyrenaicus* from Almargem when compared to the other species ($p<0.01$), while for CLOCKB *S. carolitertii* and *S. pyrenaicus* from Tagus display lower values when compared to the southern species ($p<0.01$). Aliphatic index also displayed significant changes in both proteins (KW; $p<0.001$). For CLOCKA *S. aradensis* and *S. pyrenaicus* from Almargem showed lower aliphatic index ($p<0.01$), and for CLOCKB changes were detected between *S. torgalensis*, which presented the higher value for this parameter, and all the other species ($p<0.01$). *S. aradensis* and *S. pyrenaicus* from Almargem presented intermediate values, slightly higher than the *S. carolitertii* and *S. pyrenaicus* from Tagus.

Comparing PTM sites for CLOCK proteins modifications were only for CLOCKA (Table S15). These modifications are related to phosphorylation sites: *S. pyrenaicus* from Almargem and *S. aradensis* both present an extra Protein kinase C phosphorylation site.

A structural analysis revealed the presence of three domains: two Period-Arnt-Sim (PAS fold and PAS_11) domain and the basic helix-loop-helix (bHLH). The bHLH is a protein structural motif that characterizes one of the largest families of dimerizing transcription factors and consists in a DNA-binding region. For CLOCKA most sites under negative selection are not inside the PAS domains, but in the contiguous C-terminal region of the protein.

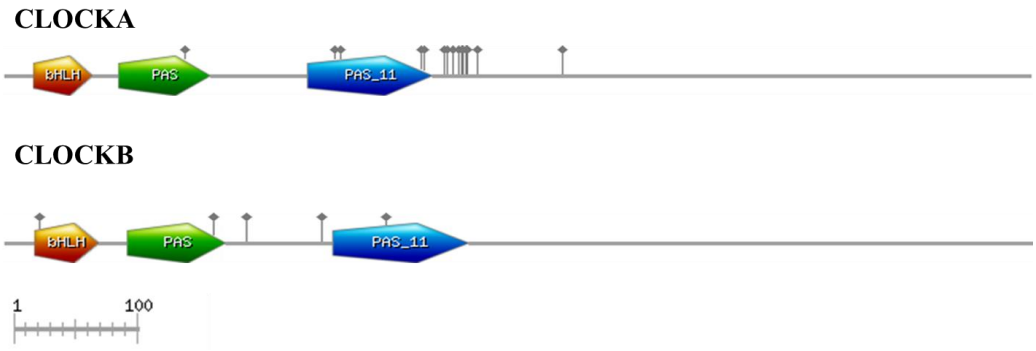

Figure 3.6. Schematic diagrams of the full length *Squalius* A) CLOCKA, and B) CLOCKB proteins. In orange is represented the basic helix-loop-helix (bHLH) motif, in green is represented the PAS fold domain and in blue is represented the Period-Arnt-Sim (PAS_11). Dots in grey represent sites under negative selection.

Protein-protein interactions for CLOCK proteins revealed only interactions with other circadian-related protein here studied (Table S16) and with CRY5 protein already mentioned for PPIs in PER family.

3.4. Evolution and characterisation of BMAL Family

For BMAL family we found three genes (Table 3.9) belonging to 2 different clades according to phylogenetic analysis (Figure 3.7). BMAL1 clade encompasses two protein-coding genes (*bmal1a* and *bmal1b*) previously identified in *D. rerio* and other fish species (Wang, 2009). BMAL2 clade is composed by a single protein-coding gene, *bmal2*, also previously characterized (Wang 2009).

Table 3.9. *Bmal* genes identified with respective annotations obtained in functional annotation analysis. ENA accession numbers are for *Squalius* sequences obtained by Sanger sequencing in this work

		Gene ontology annotation		
Gene symbol	**Gene name**	**Biological process**	**Molecular function**	**ENA accession numbers**
bmal1a (*arntl1a*)	aryl hydrocarbon receptor nuclear translocator-like 1a	· circadian system · photoperiodism · negative regulation of transcription · positive regulation of transcription · response to light stimulus	· DNA binding · protein binding · protein dimerization activity	LS999645 to LS999649
bmal1b (*arntl1b*)	aryl hydrocarbon receptor nuclear translocator-like 1b	· photoperiodism · negative regulation of transcription · positive regulation of transcription · regulation of gene expression	· DNA binding · chromatin binding · protein binding · protein dimerization activity	LS999640 to LS999644
bmal2 (*arntl2*)	aryl hydrocarbon receptor nuclear translocator-like 2	· circadian rhythm · photoperiodism · regulation of transcription · response to light stimulus	· DNA binding transcription factor activity · protein binding · protein dimerization activity	LS999568 to LS999572

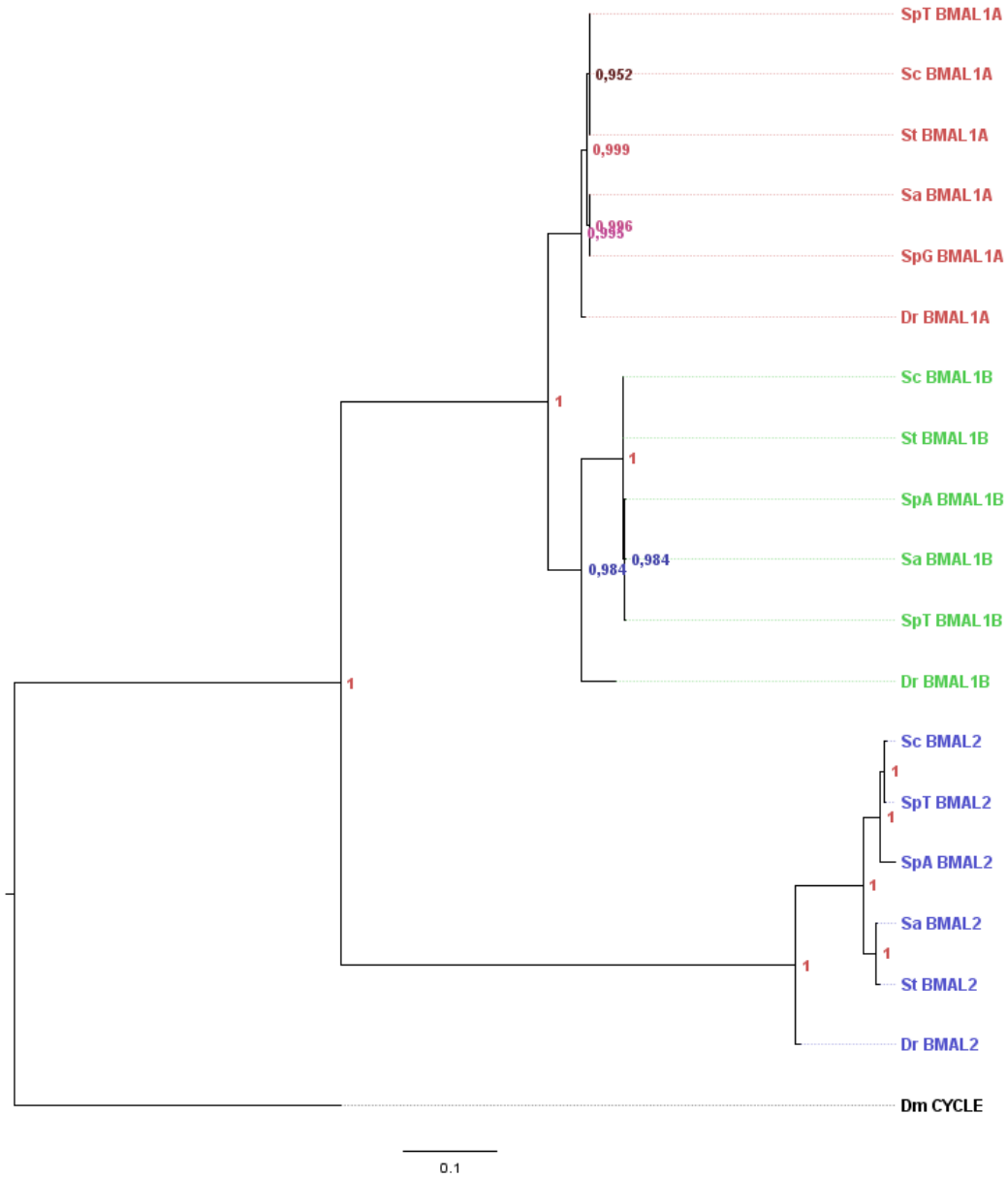

Figure 3.7. A phylogenetic tree constructed by the Bayesian Inference method for BMAL proteins with fly CYCLE protein as outgroup using JTT substitution model (Jones et al. 1992) using a discrete Gamma distribution (+G) with 3 rate categories. Values on branch nodes represent Bayesian posterior probabilities. Sc, *Squalius carolitertii*; SpT, *Squalius pyrenaicus* (Tagus population); St, Squalius torgalensis; Sa, Squalius aradensis; SpA, Squalius pyrenaicus (Almargem population); Dr, *Danio rerio*; Dm, *Drosophila melanogaster*.

Among BMAL genes, *bmal2* displayed signals of gene-wide positive (Table 3.10) but no branch-specific signals of positive selection were detected for this gene (Table 3.11). An individual site, also in *bmal2*, was found to have been subjected to episodic positive selection (Table S7). This site corresponds to a non-conservative mutation from and arginine to a proline (R20P) in *S. torgalensis* that could impact the structure of the protein, since it is well known that the presence of a proline destabilizes local structural arrangements, specifically helix geometry (Barlow & Thornton 1988). On the other hand, all *bmal* genes were found to have several sites under pervasive negative selection, specifically *bmal2* with 13 sites found (Table S17).

Table 3.10. Summary of gene-wide positive selection analysis in *bmal* genes using the BUSTED method implemented in Datamonkey webserver. A threshold of 0.1 was used for statistical significance. Rows shaded in grey correspond to genes whose test for positive selection was statically significant.

Gene	Model	Parameters	Likelihood (lnL)	AICc	LRT	P-value	ω1	ω2	ω3
bmal1a	Unconstrained model	22	-2617.3	5279.1	0	1.000	0.04 (85.93%)	0.34 (14.07%)	1.11 (0.00%)
	Constrained model	21	-2617.3	5277			0.04 (85.60%)	0.33 (14.40%)	1.00 (0.00%)
bmal1b	Unconstrained model	22	-2594.1	5232.8	0	1.000	0.01 (86.43%)	0.25 (13.57%)	1.01 (0.00%)
	Constrained model	21	-2594.1	5230.8			0.01 (86.09%)	0.24 (13.91%)	1.00 (0.00%)
bmal2	Unconstrained model	26	-2934.2	5920.8	-27	0.000	0.08 (7.00%)	0.17 (90.58%)	38.02 (2.42%)
	Constrained model	25	-2947.7	5945.8			0.00 (59.15%)	0.29 (0.00%)	1.00 (40.85%)

Table 3.11. Summary of branch-site positive selection analysis in *clock* genes using the aBSREL method implemented in Datamonkey webserver. A threshold of 0.1 was used for statistical significance. Rows shaded in grey correspond results whose test for positive selection was statically significant.

Protein	Branch	Optimized branch length	LRT	P-value	ω distribution over sites
BMAL2	*S.carolitertii*	0.002	0.000	1.000	ω1 = 0.164 (100%)
	S.pyrenaicus (Tagus)	0.000	0.000	1.000	ω1 = 1.00 (100%)
	S.torgalensis	0.001	1.134	1.000	ω1 = 10000000000 (100%)
	S.aradensis	0.000	0.000	1.000	ω1 = 1.00 (100%)
	S. pyrenaicus (Almargem)	0.005	1.376	1.000	ω1 = 3.00 (100%)

Analysing physicochemical parameters (Figures S13 – S15), changes were found for pI in BMAL2 proteins (KW; $p<0.001$) between *S. torgalensis* and *S. aradensis* and the other species ($p<0.01$). For instability we found differences in all three BMAL proteins (KW; $p<0.001$). For BMAL1A, *S. aradensis* and *S. pyrenaicus* from Almargem protein present lower instability when compared with all the other populations ($p<0.05$). BMAL1B protein from *S. torgalensis* and *S. carolitertii* presents lower values of instability compared to the remaining populations ($p<0.01$). BMAL2 presents changes in instability between all populations. For aliphatic index we detected changes in both BMAL1A and BMAL2 (KW; $p<0.001$). For BMAL1A, changes were detected between *S. aradensis* and *S. pyrenaicus* from Almargem and the other populations ($p<0.01$), and, for BMAL2 *S. aradensis* and *S. torgalensis* present for this protein lower values of aliphatic index compared to the other species ($p<0.01$), while *S. carolitertii* and *S. pyrenaicus* from Tagus present higher values ($p<0.01$). No significant changes for BMAL1B aliphatic index were detected between species (KW; $p=0.059$).

Comparing PTM sites for BMAL proteins, we only detected modifications between species for BMAL2 (Table S18). These modifications are related to phosphorylation sites: *S. torgalensis* and *S. aradensis* both present an extra Protein kinase C phosphorylation site, while the same species lack one Casein kinase II phosphorylation site when compared to the other species.

A structural analysis revealed the presence of three regions common to both BMAL proteins: two Period-Arnt-Sim (PAS fold and PAS_11) domain and the basic helix-loop-helix (bHLH) (Figure 3.8). For BMAL2 most sites under negative selection are in PAS domains, showing again the importance of these dimerization domains for protein function.

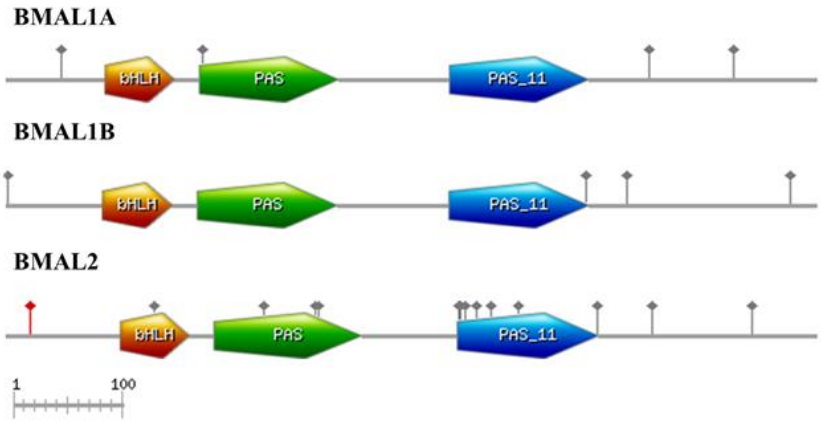

Figure 3.8. Schematic diagrams of the full length *Squalius* BMAL1A, BMAL1B, and BMAL2 proteins. In orange is represented the basic helix-loop-helix (bHLH) motif, in green is represented the PAS fold domain and in blue is represented the Period-Arnt-Sim (PAS_11). Red dots represent sites under episodic positive selection and dots in grey represent sites under negative selection.

Protein-protein interactions for BMAL proteins were found with other circadian-related proteins also studied here (Table S19) and for BMAL1A and BMAL2 interaction was found with NFIL3-5, previously mentioned in CRY family. Similarly, PER and CLOCK proteins were found to interact with CRY5 protein.

3.5. Evolution and characterisation of TIMELESS Protein

Timeless was identified as a single-copy gene (Table 3.12) and phylogenetic analysis of TIMELESS protein revealed all the species falling with a well-supported clade (Figure 3.9).

Table 3.12. *Timeless* gene identified with respective annotations obtained in functional annotation analysis. ENA accession numbers are for *Squalius* sequences obtained by Sanger sequencing in this work

		Gene ontology annotation		
Gene symbol	**Gene name**	**Biological process**	**Molecular function**	**ENA accession numbers**
timeless	timeless circadian clock	· DNA repair · DNA replication checkpoint · replication fork arrest and protection · circadian rhythm · detection of abiotic stimulus	· protein dimerization activity · protein binding · DNA binding	LS999573 to LS999577

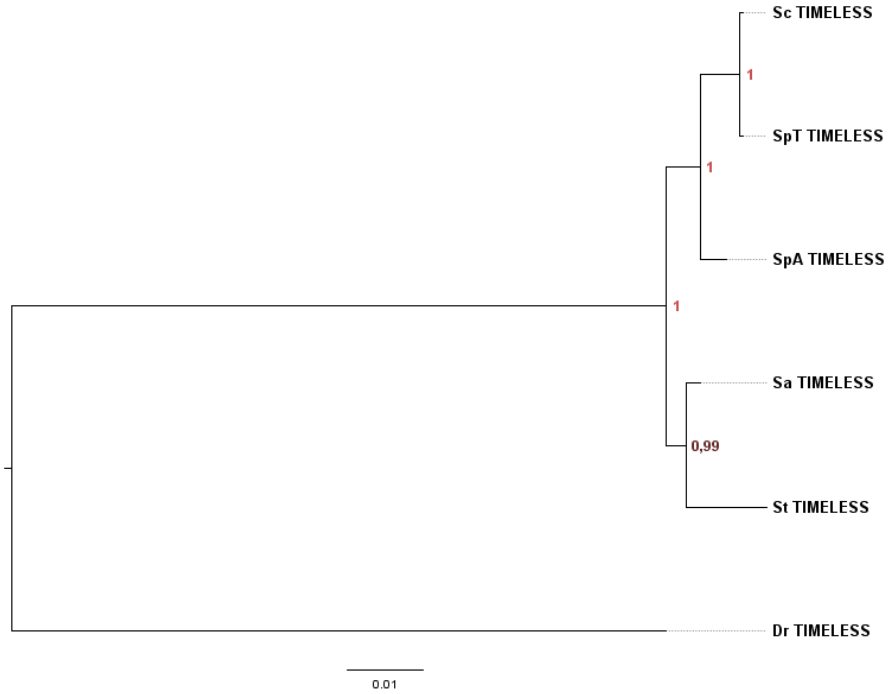

Figure 3.9. A phylogenetic tree constructed by the Bayesian Inference method for TIMELESS protein with *D. rerio* TIMELESS protein as outgroup and based on the JTT substitution model (Jones et al. 1992) with empirical amino acid frequency (+F). Values on branch nodes represent Bayesian posterior probabilities. Sc, *Squalius carolitertii*; SpT, *Squalius pyrenaicus* (Tagus population); St, *Squalius torgalensis*; Sa, *Squalius aradensis*; SpA, *Squalius pyrenaicus* (Almargem population); Dr, *Danio rerio*.

Timeless displayed gene-wide signals of positive selection (Table 3.13) whose branch analysis revealed to be in *S. torgalensis, S. pyrenaicus* (Tagus) and *S. carolitertii* (Table 3.14). It was also identified a site to be under pervasive positive selection associated to a non-conservative mutation (L360N) (Table S7) present in the species whose branches experienced positive selection. This mutation from leucine (nonpolar amino acid) to asparagine (polar amino acid) can impact either the function and protein structure, since the introduction of a polar amino acid could induce a disruption in protein core structure.

Table 3.13. Summary of gene-wide positive selection analysis in *timeless* gene using the BUSTED method implemented in Datamonkey webserver. A threshold of 0.1 was used for statistical significance. Rows shaded in grey correspond to genes whose test for positive selection was statically significant.

Gene	Model	Parameters	Likelihood (lnL)	AICc	LRT	P-value	ω1	ω2	ω3
timeless	Unconstrained model	24	-5439.7	10927.6	-9.6	0.008	0.00 (69.31%)	0.00 (25.90%)	28.06 (4.79%)
	Constrained model	23	-5444.5	10935.2			0.01 (0.00%)	0.64 (0.00%)	1.00 (100.00%)

Table 3.14. Summary of branch-site positive selection analysis in *timeless* gene using the aBSREL method implemented in Datamonkey webserver. A threshold of 0.1 was used for statistical significance. Rows shaded in grey correspond results whose test for positive selection was statically significant.

Protein	Branch	Optimized branch length	LRT	P-value	ω distribution over sites
TIMELESS	*S.carolitertii / S.pyrenaicus (Tagus)*	0.002	6.149	0.083	ω1 = 0.00 (99%); ω2 = 60.9 (1.5%)
	S.torgalensis	0.004	5.520	0.091	ω1 = 0.186 (95%); ω2 = 47.1 (4.5%)
	S.aradensis	0.001	0.564	0.634	ω1 = 10000000000 (100%)
	S.pyrenaicus (Almargem)	0.002	0.600	0.930	ω1 = 10000000000 (100%)

TIMELESS physicochemical parameters analysed presented significant changes (KW; $p<0.001$) (Figure S16). A small increase in pI was detected in *S. aradensis* and *S. torgalensis* ($p<0.05$); TIMELESS from southern populations, specifically *S. torgalensis*, shows higher instability compared to TIMELESS from *S. carolitertii* and *S. pyrenaicus* from Tagus ($p<0.01$). TIMELESS from *S. torgalensis* presents a reduced value of aliphatic index ($p<0.01$) compared to all the other species. Furthermore, TIMELESS from *S. pyrenaicus* from Almargem presents a slightly higher aliphatic index compared to all the other populations ($p<0.01$). Aliphatic index for *S. aradensis* TIMELESS presents a small reduction compared to the northern populations ($p<0.05$), but a higher value than TIMELESS from its sister species, *S. torgalensis* ($p<0.001$).

TIMELESS protein also presents alterations in post-translational modification sites, specifically phosphorylation sites where TIMELESS from *S. aradensis* and *S. torgalensis* presents an extra protein kinase C phosphorylation site, and *S. aradensis* and *S. pyrenaicus* from Almargem present an extra casein kinase II phosphorylation site (Table S21).

TIMELESS protein presents two main domains: the TIMELESS domain and Timeless protein C terminal region (Figure 3.10). However, little is known about the function of these domains in vertebrates.

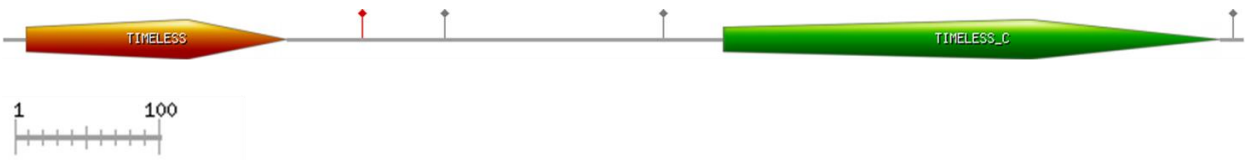

Figure 3.10. Schematic diagrams of the full length *Squalius* TIMELESS protein. In orange is represented the TIMELESS domain and in green is represented the Timeless protein C terminal region. Red dots represent sites under episodic positive selection and dots in grey represent sites under negative selection.

Protein-protein interactions in TIMELESS protein were poorly related to the circadian rhythm, as this protein only interacts with CRY1AA and CRY2 (Table S22). However, several interactions were found between TIMELESS and proteins related to DNA replication and cell cycle regulation, namely DNA Helicases (MCM2, MCM9, MCM4 and MCM3L), DNA polymerase I (POLA1), and checkpoint kinase 1 (CHEK1) (Table S22).

4. Discussion

The circadian system involves of a complex network of regulatory pathways linking several genes and proteins. Fifteen genes belonging to four main gene families related to the circadian systems and a single-copy gene also putatively related with circadian maintenance were here identified in four related freshwater fish species of *Squalius* genus from the western Iberian Peninsula. Except for CLOCK family for which *clock2* gene was not identified in available transcriptomes of *S. carolitertii* and *S. torgalensis* (Jesus et al. 2015), the number of duplicates in each gene family matched previous descriptions for zebrafish (Wang 2008a, 2008b, 2009; Liu et al. 2015). After gene sequencing, protein sequences were predicted and characterised at different levels. A phylogenetic analysis of each family based on predicted protein sequences for *Squalius* species supported the evolutionary history of duplications reported for other fish species (Toloza-Villalobos et al. 2015). Moreover, additional patters of adaptation were observed for a set of genes in the different populations, most probably reflecting differences in light and temperature to which these species are subjected.

Diversification in functions of circadian-related proteins may arise to optimize important biological processes in space and time, synchronised with the circadian oscillation. Characterisation of protein-protein interactions (PPI) revealed that the studied proteins retain functions in the circadian system common to other vertebrates. Nevertheless, most PER and CRY proteins were also found to interact with BHLHe41. CRY2 was found to interact with HSF2 as well. HSF2 is a transcription factor involved in activation of HSP expression in conditions of thermal stress. This result supports other findings that connected the circadian system with the thermal-stress response via heat-shock proteins (HSP) (Tamaru et al. 2011; Chappuis et al. 2013; Jerônimo et al. 2017). On the other hand, BHLHe41 is a protein associated with biochemical pathways activated in response to cold, as previously observed in cold shock responses in zebrafish larvae (Hung et al. 2016). Thus, it is here reinforced the important interaction between circadian system and temperature, as previously reported (e.g. Lahiri et al. 2005; Chappuis et al. 2013; Jerônimo et al. 2017). In addition, BMAL1A, BMAL2 and CRY3 were found to interact with NFIL3-5 (nuclear factor, interleukin 3-regulates, member 5), a protein involved in immune response, specifically in the activation of transcription from the interleukin-3 promoter in T-cells (Zhang et al. 1995). It was previously revealed for zebrafish that *nfil3-5* gene is rhythmically expressed in a light-dependent way, an indication this gene may be coordinated with the circadian rhythmicity, reinforcing the link between the immune and circadian systems already established (Li et al. 2015). Furthermore, PPI allowed to provide some insights on the evolution of *Squalius* TIMELESS, namely in providing some information concerning its function. Apparently, TIMELESS protein retains domains homologs to *Drosophila* TIMELESS which allow the protein to interact with CRY proteins, namely CRY1AA and CRY2. Therefore, PPI supported the findings of some works that present the hypothesis that *timeless* gene has a putative circadian-related function by being an alternative to PER proteins in the negative loop of the system (Barnes et al. 2003). However, it is shown here that TIMELESS interacts mainly with cell cycle-related proteins, specifically helicases, DNA polymerases, and checkpoint kinases, enzymes related to DNA replication. This is not surprising as mouse and human homologs were already implicated in processes involved with cell division and development (Yoshizawa-Sugata & Masai 2007). This points to a specialization of TIMELESS towards circadian dependant regulation of cell cycle as also suggested for other vertebrates (Yoshizawa-Sugata & Masai 2007).

4.1. Evolution of circadian-related proteins in western Iberian *Squalius*

To better understand patters of molecular evolution of circadian-related genes and adaptation of circadian rhythms to different conditions of light and temperature in the Portuguese *Squalius* species phylogenetic relationships were used together with signatures of selection and protein analysis.

In previous studies, a multi-locus phylogenetic analysis of these species with nuclear markers recovered two main clusters: i) *S. aradensis* and *S. torgalensis*; ii) *S. pyrenaicus* from Tagus clustering with *S. carolitertii*. Southern populations of *S. pyrenaicus* (Almargem and Guadiana) rendered paraphyletic to *S. carolitertii*, suggesting that the Central and Southern populations of *S. pyrenaicus* could be considered as distinct species (Waap et al. 2011; Coelho et al. unpublished data). Nevertheless, this debate is out of the scope of this study.

For CRY1BA, PER1A, PER2, BMAL2 and TIMELESS phylogenetic tree recovered the topology from the species tree mentioned above. For the mentioned proteins, signatures of gene-wide positive selection were found for the coding genes *per1a, per2* and *bmal2*, and were further confirmed by protein analysis. For PER1A, the mutation T254A in *S. carolitertii* was detected to be under positive selection, while for *per2,* signatures of selection were detected in southern sister species *S. aradensis* and *S. torgalensis*. For *bmal2* only gene-wide positive selection was detected. For *cry1ba* gene no signals of positive selection were detected with any test, but protein analysis reveals important functional modifications in CRY1BA for circadian system.

In PER1A from *S. carolitertii*, the mutation T254A adds to the protein an extra aliphatic amino acid leading to an increase in aliphatic index of the protein, i.e. in its thermostability, as effectively predicted in protein analysis. Moreover, higher protein thermostability was found for *S. carolitertii* and both *S. pyrenaicus* populations (Tagus and Almargem) in CRY1BA, PER1A and BMAL2, resulting in better aptitude of these proteins to deal with thermal instability. Additionally, post-translational modifications (PTM) patterns for these proteins seem to corroborate observed trends in protein parameters, as for all of them, *S. carolitertii* and both populations of *S. pyrenaicus* (Tagus and Almargem) seem to have PTM sites conserved, specifically protein kinase C phosphorylation sites and cAMP- and cGMP-dependent protein kinase phosphorylation sites. Phosphorylation is one of the most important PTM in circadian regulation (Reischl & Kramer 2011), so these sites must reflect selective pressures for increased regulatory demands. Moreover, in CRY1BA and BMAL2 some of the sites under negative selection are in codons coding for serine, threonine and tyrosine, amino acids involved in phosphorylation. Thus, it seems that selective pressure is keeping these sites conserved in the protein, as mutation changing these amino acids are negatively selected. N-glycosylation is another important PTM in intracellular proteins, and has been shown to have important roles in protein folding and stabilisation, which is positively correlated with the degree of glycosylation (Shental-Bechor & Levy 2008). N-glycosylation sites were detected to be highly conserved in all proteins, but in PER1A, *S. carolitertii* and both populations of *S. pyrenaicus* present extra N-glycosylation sites, and increased number these sites in may reflect higher regulatory demands towards the process of folding.

Conversely, PER2 from *S. pyrenaicus* (Almargem) seems to be slightly differentiated from the other population of *S. pyrenaicus* (Tagus) and *S. carolitertii*. This differentiation also supports the paraphyly of *S. pyrenaicus* to *S. carolitertii* (Waap et al. 2011; Coelho et al. unpublished data) and seems to be the result of a point mutation (T402D) which was detected to be positively selected. This mutation leads to changes in protein features, namely in PTM sites, as in this southern population PER2 lost a threonine, and consequently a phosphorylation site. Hyperphosphorylation was reported to be a signal for PER2 degradation, so mutations that caused a decreasing the number of

phosphorylation sites led to an increase protein stability in the cell, and consequently, *in vivo* half-life (Edery et al. 1994).

In addition, for *per2*, signatures of selection were detected in both southwestern sister species *S. aradensis* and *S. torgalensis* and confirmed by subsequent protein analysis that revealed similar functional features in PER2 from these species. Such pattern of adaptation in these populations may be directly related to PER2 functions, as this protein was reported to have a role in response to UV radiation. PER2 may then have an enhanced importance in regions with increased intensity of radiation. In fact, presently these regions experience higher solar irradiance when compared to the northern region (IPMA [ipma.pt (Portuguese Institute for Sea and Atmosphere); accessed in June 2018]; data not showed). In addition, PPI for PER2 revealed an interaction with CRY-DASH protein, a cryptochrome-related protein strongly implied in UV-induced DNA damage repair (Daiyasu et al. 2004; Selby & Sancar 2006).

Finally, *timeless* gene was found to be evolving under positive selection, specifically in *S. carolitertii*, *S. pyrenaicus* from Tagus and *S. torgalensis.* In *S. torgalensis* a decrease in protein thermostability that can be correlated with the positively selected mutation L360N. However, this mutation introduces an asparagine in the protein, an amino acid able to establish stronger hydrogen bonds compared to the amino acid substituted. Therefore, this mutation may have a positive impact in the stabilisation of TIMELESS three-dimensional structure. Moreover, this strong positive selection in three branches of the phylogeny may be an indicator of additional diversification of this protein in these species, and requires further investigation.

Concluding, the observed patterns for CRY1BA PER1A, PER2 and BMAL2 and TIMELESS proteins for the different species can be related with the geomorphological history of the basins (e.g. Brito et al. 1997; Almada & Sousa-Santos 2010; Coelho et al. unpublished data), but adaptive patterns described for PER2 can be a contribution to explain the paraphyly in *S. pyrenaicus*.

4.2. Evolution of circadian-related proteins according the N-S distribution

In the present, the distribution of *Squalius* in Portugal follows a latitudinal gradient with specific environmental features. *S. carolitertii* and Tagus population of *S. pyrenaicus* inhabit colder rivers in the North of Portugal, exposed to a lower number of daylight hours. In contrast, *S. torgalensis*, *S. aradensis* and Almargem population of *S. pyrenaicus* inhabit southern rivers with higher water temperature and increased number of daylight hours (Table S1). Additionally, water in southern basis tends to be more alkaline than in northern basins. Fish, as ectothermic, depend strictly from the environmental conditions and need to have strategies to overcome changes in environmental conditions.

For CRY1AA, CRY1BB, CRY3, PER3 and CLOCKB, phylogenetic analysis of these protein revealed two main clustering groups: i) *S. carolitertii* and *S. pyrenaicus* from Tagus; and ii) *S. torgalensis*, *S. aradensis* and *S. pyrenaicus* from Almargem. This topology is incongruent with the species tree but is congruent with the N-S distribution pattern of these species relatively to environmental conditions of light and temperature. Thus, this dichotomy between the two main groups can be a consequence of adaptations related with the similar environmental conditions experienced by each group of populations.

Both PER3 and CLOCKB presented differences in pI. In the northern populations studied (*S. carolitertii* and *S. pyrenaicus* from Tagus), the mutation K429Y in PER3 was found to be under positive selection in both species, and this mutation is responsible for replacing a lysine for a tyrosine, a basic for a neutral amino acid, leading to a decrease of PER3 pI in the mentioned populations. In CLOCKB, a single mutation (F508R) was found to be responsible for this change in pI, but no

signatures of positive selection were found acting on this site. This mutation occurs in *S. torgalensis*, *S. aradensis* and *S. pyrenaicus* from Almargem, and changes a phenylalanine to an arginine, leading to a change from a neutral to a basic amino acid, which results to an increase in pI. The increase in pI was not expected in these species, but this pattern may be related to specific uncharacterised functions of these proteins, and so, requires further investigation.

For CRY1AA, it was not possible to establish supported phylogenetic relationships, as phylogenetic relationships in external branches appear as a polytomy. Even so, CRY1AA from *S. torgalensis* appears to be functionally differentiated from the other species. For CRY1AA, the physicochemical parameter differentiating *S. torgalensis* from other species is pI, which can be correlated with the mutation K275Q, a site under positive selection. This mutation is responsible for changing a lysine to a glutamine, i.e. an uncharged to a positively charged amino acid with acid properties, leading to a decrease in the overall charge of the protein, and consequently in pI. Interestingly, in this protein the FAD binding domain seems to be highly conserved with several sites located in this domain consistently under negative selection. Previous studies with *S. torgalensis* and *S. carolitertii* already revealed that expression of *cry1aa* is strongly affected either by water temperature and pH in *S. carolitertii,* while in *S. torgalensis* expression of these genes seemed to only be affected by pH (Jesus et al. 2017). Therefore, the mentioned work already provided some indications that CRY1AA from *S. torgalensis* may show specific adaptations to deal with high temperatures (Jesus et al. 2017). The decrease in pI can also explain the changes in *cry1aa* expression in conditions of water acidification, since pI is highly related to protein solubility and changes in this property could lead to protein precipitation reducing its activity to critical levels (Kiraga et al. 2007; Khaldi & Shields 2011).

This pattern of adaptation observed is not surprising as CRY1AA was pointed to be a key agent in the core of circadian system together with PER2 (Tamai et al. 2007). Taking another look at PER2 protein, the positively selected mutation T402D is in PAS-3 domain. This mutation changes a threonine to an aspartate, which chemically results in the change of an uncharged to a charged amino acid. PAS-3 domain is an important domain for dimerization of PER2 with CRY1AA (Ponting & Aravind 1997; Tamai et al. 2007), and charged amino acids are typically more prone to form ionic interactions with other amino acids, and this interaction is mainly responsible for dimer formation (Bhagavan & Ha 2011). Thus, an extra charge in this region can be a positive factor for PER:CRY dimer stability. PER2 was also the protein to present the higher number of sites under negative selection, mostly inside functional domains, reinforcing the importance of this protein in maintaining the core loop in proper conditions.

However, for *S. torgalensis,* other phylogenetic relationships were established. In some cases, this species clusters with northern populations *S. carolitertii* and *S. pyrenaicus* from Tagus, an unexpected pattern due to either the evolutionary history of these species, or the different environmental conditions experienced in the North and in the South. This unexpected clustering was observed in PER1B, BMAL1A and BMAL1B phylogenies. For PER1B, signatures of positive selection are present in *S. torgalensis* showing a possible independent evolution in this species. Protein analysis revealed PER1B from *S. torgalensis* seems to be thermostable than PER1B from all the other species. This increase in thermostability is most probably related to an increase in aliphatic index caused by the mutation T1256V in *S. torgalensis*, found to be positively selected. This protein has been implied as one of the most important proteins in integrating temperature and light cues within the circadian system (Lahiri et al. 2005). Even though the number of light hours per day and water temperature experienced in Mira basin are similar to those experienced in other southern basis, it has specific environmental characteristics. Data from SNIRH revealed Mira basin has differences in annual temperature variation compared to Arade or Almargem (8.7ºC for Mira vs 10.5ºC for Arade). Hence,

water temperature in Mira basin seems have lower fluctuations during the year compared to other southern basins. This may be the result of a relative overshadowing of rivers from Mira basin compared to the other southern basins caused by the presence of riparian vegetation in some ranges of Mira basin. Density of riparian vegetation was shown to influence several aspects in the dynamics of the slow-flowing rivers, specifically water temperature, but also the irradiance able to get to the water (Garner et al. 2017; Trimmel et al. 2018). Altogether, these conditions impose *S. torgalensis* harsher conditions in a regular basis demanding for optimised pathways of integrating temperature within the circadian system, where PER1B displays an important role. Hence, *S. torgalensis* may be exhibiting signals of local adaptation for an improved pathway of temperature-circadian regulation that should be further explored. Furthermore, it is interesting to denote that in *per1b*, several sites dispersed across the protein in codons coding for aliphatic amino acids are negatively selected underpinning higher demands for protein thermostability.

4.2.1. Adaptive convergence in *Squalius aradensis* and *Squalius pyrenaicus* from Almargem basin

Phylogenetic trees revealed in 7 out of 16 proteins (CRY1BB, CRY3, PER3, CLOCKA, CLOCKB, BMAL1A and BMAL1B) show a well-supported cluster holding *S. aradensis* and *S. pyrenaicus* from Almargem, incongruent with the species tree. This suggests a pattern of a strong evolutionary convergence between these populations. Almargem and Arade are two basins in the south of Portugal extremely influenced by Mediterranean climate. Moreover, the environmental conditions experienced in these basins are extremely similar, including average water temperature, water pH and photoperiod. So, this pattern of convergence may be a consequence of similar selective pressures imposed for these similar environmental conditions in the two basins. Similar patterns of evolutionary convergence have been commonly described for other fish species as a consequence of selective pressures imposed by similar environmental conditions (e.g. Protas et al. 2006; Muschick et al. 2012; Nath et al. 2013; Alter et al. 2015; Passow et al. 2017).

For CRY1BB, a poorly studied protein in other fish species, PPI revealed possible interactions with other circadian-related proteins but did not display any signals of post-duplication diversification. Also, its function seems to be redundant, as other CRY proteins displayed the same pattern of protein-protein interactions. Therefore, this protein may be evolving without strong constraints or under neutrality. For BMAL1A and BMAL1B, there is not enough evidence for signatures of selection on respective coding genes. However, protein analysis exposed some patterns consistent with the hypothesis of convergence between *S. aradensis* and *S. pyrenaicus* from Almargem. Specifically, for BMAL1A, higher thermostability was detected in *S. aradensis* and *S. pyrenaicus* from Almargem, which can be explained with the similar environmental conditions experienced by these populations.

Concerning the genes coding for proteins that revealed signals of convergent evolution, *per3*, *clocka* and *clockb* present signatures of gene-wide positive selection. In *clocka* these signatures were found in *S. aradensis* and *S. pyrenaicus* from Almargem, but in *per3* signatures were only found in *S. pyrenaicus* from Almargem. Nevertheless, protein analysis revealed all these proteins are functionally similar in *S. aradensis* and *S. pyrenaicus* from Almargem. Altogether, signatures of selection and protein analysis reinforced the possible convergent evolution of these proteins in *S. aradensis* and *S. pyrenaicus* from Almargem. CRY3, PER3 and CLOCKB exhibited higher thermostability in these populations that may be related to the more similar higher temperatures experienced in their basins, but PPI patterns of these proteins revealed they are exclusively related with circadian system regulation. Hence, they seem to be not involved in any other biochemical function within the circadian system (e.g. entrainment of circadian system by temperature). Proteins related with responses to temperature stimuli are more thermostable per se due to its intrinsic function to deal with thermal

stress or by interacting with other thermostable proteins, but proteins lacking these interactions are strictly dependent on mutations that affect positively their physicochemical properties (Panja et al. 2015). Hence, mutations that increase aliphatic index of the protein may be positively selected in environments as those experienced in Arade and Almargem, where water temperatures are higher. Moreover, we observed several sites under negative selection in CRY3 and PER3 that can be responsible for increasing thermostability, as they occur in codons that encode for aliphatic amino acids. In CRY3, these sites seem to be particularly relevant as they are all located inside functional domains of the protein. In addition, PER3 and CLOCKA also displayed patterns of PTM congruent with the hypothesis of convergence, as the number of protein kinase c phosphorylation sites is similar in both *S. aradensis* and *S. pyrenaicus* from Almargem. Altogether, these results integrating signatures of selection and protein characterisation at their different levels exposed the pattern of convergence between *S. aradensis* and *S. pyrenaicus* from Almargem. This evolutionary convergence can be explained by the more similar environmental conditions experienced in Arade and Almargem, basins the inhabit.

5. Final Remarks

Sequence-based gene and protein analysis allowed to provide here the first insights on the evolution of circadian system related gene families in *Squalius* from the western Iberian Peninsula. For all studied proteins, characterisation of protein-protein interactions displayed signals of post-duplication diversification in circadian-related proteins from CRY, PER and BMAL families in *Squalius* species. It was also possible to provide some insights on TIMELESS function, a protein whose circadian function has been under debate in the past few years. A putative function within the circadian system was exposed, but most of TIMELESS function appears to be related with cell cycle regulation. However, this observation raises some open questions concerning TIMELESS function that require further experimental characterisation, specifically to clarify the mechanisms by which this protein performs its function.

Altogether, phylogenetic studies, signatures of selection and protein analysis revealed the evolutionary history of circadian-related proteins in the western Iberian *Squalius* species. Phylogenies recovered for some proteins (CRY1BA PER1A, PER2, BMAL2 and TIMELESS) reflect the evolutionary history of these species recovered by other nuclear genes (Waap et al. 2010). Furthermore, for PER2 the higher observed differentiation between *S. pyrenaicus* from Almargem compared to *S. carolitertii* and *S. pyrenaicus* from Tagus support the paraphyly of *S. pyrenaicus* towards *S. carolitertii*.

For some proteins (CRY1AA, CRY1BB, CRY3, PER3 and CLOCKB), the different species evidence adaptive patterns that seem to reflect the N-S gradient of light and temperature. Moreover, for *S. torgalensis* a pattern of local adaptation was found for two proteins (PER1B and CRY1AA), which appear to have evolved independently in this species as a response to specific environmental conditions experienced in Mira basin.

An interesting result of this study was a well-supported convergence between the southern populations of *S. aradensis* and *S. pyrenaicus* from Almargem for proteins CRY1BB, CRY3, PER3, CLOCKA, CLOCKB, BMAL1A and BMAL1B. Signatures of selection supported this convergence in *clocka*, but for other genes, this pattern of convergence was only exposed after protein characterisation. For these proteins, populations show similar functional features, which is explained by the similar environmental conditions that these populations experience in their habitats.

In addition, it was possible to show that temperature may have strongly influenced the evolution of circadian-related proteins in *Squalius* species, as most proteins exhibit a strong response in thermostability.

Here, by complementing selection tests based on dN/dS with protein analysis it was possible to characterize and detect selection. Alone, tests based on dN/dS were not sufficient, suggesting that these tests do not have suitable statistical power to detect signatures of selection. This reinforces the idea that dN/dS tests should be complemented with other layers of analysis, following and integrative approach including a sequence-based protein characterisation at their different structural levels of organisation.

Finally, further investigation is required, specifically to possibly identify *clock2* gene using the recently published genome of the freshwater fish *Leuciscus waleckii*, more closely related to the *Squalius* genus than *Danio rerio*, as well as new transcriptomes of *Squalius*, recently analysed in our research group (unpublished data).

References

Abascal F, Zardoya R, Posada D. 2005. ProtTest: Selection of best-fit models of protein evolution. Bioinformatics. 21:2104–2105. doi: 10.1093/bioinformatics/bti263.

Almada V, Sousa-Santos C. 2010. Comparisons of the genetic structure of Squalius populations (Teleostei, Cyprinidae) from rivers with contrasting histories, drainage areas and climatic conditions based on two molecular markers. Mol. Phylogenet. Evol. 57:924–931. doi: 10.1016/j.ympev.2010.08.015.

Alter SE, Brown B, Stiassny MLJ. 2015. Molecular phylogenetics reveals convergent evolution in lower Congo River spiny eels. BMC Evol. Biol. 15:224. doi: 10.1186/s12862-015-0507-x.

Asher G, Sassone-Corsi P. 2015. Time for food: The intimate interplay between nutrition, metabolism, and the circadian clock. Cell. 161:84–92. doi: 10.1016/j.cell.2015.03.015.

Barlow DJ, Thornton JM. 1988. Helix geometry in proteins. J. Mol. Biol. 201:601–619. doi: 10.1016/0022-2836(88)90641-9.

Barnes JW et al. 2003. Requirement of mammalian Timeless for circadian rhythmicity. Science. 302:439–442. doi: 10.1126/science.1086593.

Ben-Moshe Z et al. 2014. The light-induced transcriptome of the zebrafish pineal gland reveals complex regulation of the circadian clockwork by light. Nucleic Acids Res. 42:3750–3767. doi: 10.1093/nar/gkt1359.

Bhagavan NV, Ha C-E. 2011. Chapter 4 - Three-Dimensional Structure of Proteins. In: Essentials of Medical Biochemistry. Bhagavan, NV & Ha, C-E, editors. Academic Press: San Diego pp. 29–38. doi: 10.1016/B978-0-12-095461-2.00004-7.

Brito RM, Briolay J, Galtier N, Bouvet Y, Coelho MM. 1997. Phylogenetic relationships within genus Leuciscus (Pisces, Cyprinidae) in Portuguese fresh waters, based on mitochondrial DNA cytochrome b sequences. Mol. Phylogenet. Evol. 8:435–442. doi: 10.1006/mpev.1997.0429.

de Castro E et al. 2006. ScanProsite: Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 34. doi: 10.1093/nar/gkl124.

Cathala G et al. 1983. A Method for Isolation of Intact, Translationally Active Ribonucleic Acid. DNA. 2:329–335. doi: 10.1089/dna.1983.2.329.

Chappuis S et al. 2013. Role of the circadian clock gene Per2 in adaptation to cold temperature. Mol. Metab. 2:184–193. doi: 10.1016/j.molmet.2013.05.002.

Coelho MM, Bogutskaya NG, Rodrigues JA, Collares-Pereira MJ. 1998. Leuciscus torgalensis, and L. aradensis, two new cyprinids for Portuguese fresh waters. J. Fish Biol. 52:937–950. doi: 10.1111/j.1095-8649.1998.tb00594.x.

Daiyasu H et al. 2004. Identification of cryptochrome DASH from vertebrates. Genes Cells Devoted Mol. Cell. Mech. 9:479–495. doi: 10.1111/j.1356-9597.2004.00738.x.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 27:1164–1165. doi: 10.1093/bioinformatics/btr088.

Dawson NL et al. 2017. CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Res. 45:D289–D295. doi: 10.1093/nar/gkw1098.

Di Tommaso P et al. 2011. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res. 39:W13–W17. doi: 10.1093/nar/gkr245.

Dunlap JC. 1999. Molecular Bases for Circadian Clocks. Cell. 96:271–290. doi: 10.1016/S0092-8674(00)80566-8.

Eddy SR. 2011. Accelerated Profile HMM Searches Pearson, WR, editor. PLoS Comput. Biol. 7:e1002195. doi: 10.1371/journal.pcbi.1002195.

Eddy SR. 1998. Profile hidden Markov models. Bioinformatics. 14:755–763. doi: 10.1093/bioinformatics/14.9.755.

Edery I, Zwiebel LJ, Dembinska ME, Rosbash M. 1994. Temporal phosphorylation of the Drosophila period protein. Proc. Natl. Acad. Sci. U. S. A. 91:2260–2264.

Finn RD et al. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44:D279–D285. doi: 10.1093/nar/gkv1344.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39:W29–W37. doi: 10.1093/nar/gkr367.

Foulkes NS, Whitmore D, Vallone D, Bertolucci C. 2016. Studying the Evolution of the Vertebrate Circadian Clock. In: Genetics, Genomics and Fish Phenomics.Vol. 95 pp. 1–30. doi: 10.1016/bs.adgen.2016.05.002.

Garner G, Malcolm IA, Sadler JP, Hannah DM. 2017. The role of riparian vegetation density, channel orientation and water velocity in determining river temperature dynamics. J. Hydrol. 553:471–485. doi: 10.1016/j.jhydrol.2017.03.024.

Gasteiger E et al. 2005. Protein Identification and Analysis Tools on the ExPASy Server. Proteomics Protoc. Handb. 571–607. doi: 10.1385/1-59259-890-0:571.

Gotter AL. 2006. A Timeless debate: resolving TIM's noncircadian roles with possible clock function: NeuroReport. 17:1229–1233. doi: 10.1097/01.wnr.0000233092.90160.92.

Haft DH et al. 2013. TIGRFAMs and Genome Properties in 2013. Nucleic Acids Res. 41:D387–D395. doi: 10.1093/nar/gks1234.

Huang D et al. 2007. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol. 8:R183. doi: 10.1186/gb-2007-8-9-r183.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755. doi: 10.1093/bioinformatics/17.8.754.

Hulo N et al. 2008. The 20 years of PROSITE. Nucleic Acids Res. 36:D245–D249. doi: 10.1093/nar/gkm977.

Hung I-C, Hsiao Y-C, Sun HS, Chen T-M, Lee S-J. 2016. MicroRNAs regulate gene plasticity during cold shock in zebrafish larvae. BMC Genomics. 17. doi: 10.1186/s12864-016-3239-4.

Ishikawa T, Hirayama J, Kobayashi Y, Todo T. 2002. Zebrafish CRY represses transcription mediated by CLOCK-BMAL heterodimer without inhibiting its binding to DNA. Genes Cells. 7:1073–1086. doi: 10.1046/j.1365-2443.2002.00579.x.

Jerônimo R et al. 2017. Thermal stress in Danio rerio : a link between temperature, light, thermo-TRP channels, and clock genes. J. Therm. Biol. 68:128–138. doi: 10.1016/j.jtherbio.2017.02.009.

Jesus T et al. 2015. Data from: 'Characterization of two Iberian freshwater fish transcriptomes, Squalius carolitertii and Squalius torgalensis, living in distinct environmental conditions' in Genomic Resources Notes Accepted 1 April 2015 to 31 May 2015. Mol. Ecol. Resour. 16:377. doi: 10.5061/dryad.fm28d.

Jesus TF et al. 2017. Protein analysis and gene expression indicate differential vulnerability of Iberian fish species under a climate change scenario Rutherford, S, editor. PLOS ONE. 12:e0181325. doi: 10.1371/journal.pone.0181325.

Jesus TFF, Grosso ARR, Almeida-Val VMF, Coelho MM. 2016. Transcriptome profiling of two Iberian freshwater fish exposed to thermal stress. J. Therm. Biol. 55:54–61. doi: 10.1016/j.jtherbio.2015.11.009.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Bioinformatics. 8:275–282. doi: 10.1093/bioinformatics/8.3.275.

Khaldi N, Shields DC. 2011. Shift in the isoelectric-point of milk proteins as a consequence of adaptive divergence between the milks of mammalian species. Biol. Direct. 6:40. doi: 10.1186/1745-6150-6-40.

Kiraga J et al. 2007. The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. BMC Genomics. 8:163. doi: 10.1186/1471-2164-8-163.

Kornmann B, Schaad O, Bujard H, Takahashi JS, Schibler U. 2007. System-Driven and Oscillator-Dependent Circadian Transcription in Mice with a Conditionally Active Liver Clock. PLOS Biol. 5:e34. doi: 10.1371/journal.pbio.0050034.

Kosakovsky Pond SL et al. 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol. Biol. Evol. 28:3033–3043. doi: 10.1093/molbev/msr125.

Kosakovsky Pond SL, Frost SDW. 2005a. Datamonkey: Rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics. 21:2531–2533. doi: 10.1093/bioinformatics/bti320.

Kosakovsky Pond SL, Frost SDW. 2005b. Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. 22:1208–1222. doi: 10.1093/molbev/msi105.

Kosakovsky Pond SL, Frost SDW, Muse VS. 2005. HyPhy: Hypothesis testing using phylogenies. Bioinformatics. 21:676–679. doi: 10.1093/bioinformatics/bti079.

Lahiri K et al. 2005. Temperature Regulates Transcription in the Zebrafish Circadian Clock Stemple, D, editor. PLoS Biol. 3:e351. doi: 10.1371/journal.pbio.0030351.

Lazado CC et al. 2014. Daily Rhythmicity of Clock Gene Transcripts in Atlantic Cod Fast Skeletal Muscle Oster, H, editor. PLoS ONE. 9:e99172. doi: 10.1371/journal.pone.0099172.

Le SQ, Gascuel O. 2008. An Improved General Amino Acid Replacement Matrix. Mol. Biol. Evol. 25:1307–1320. doi: 10.1093/molbev/msn067.

Li L, Ji D, Teng L, Zhang S, Li H. 2015. Identification and expression of lypc, a novel dark-inducible member of Ly6 superfamily in zebrafish Danio rerio. Gene. 574:69–75. doi: 10.1016/j.gene.2015.07.088.

Lim C, Allada R. 2013. Emerging roles for post-transcriptional regulation in circadian clocks. Nat. Neurosci. 16:1544–1550. doi: 10.1038/nn.3543.

Lin C, Todo T. 2005. The cryptochromes. Genome Biol. 6:220. doi: 10.1186/gb-2005-6-5-220.

Liu C et al. 2015. Molecular evolution and functional divergence of zebrafish (Danio rerio) cryptochrome genes. Sci. Rep. 5:8113. doi: 10.1038/srep08113.

Liu Y, Merrow M, Loros JJ, Dunlap JC. 1998. How temperature changes reset a circadian oscillator. Science. 281:825–829. doi: 10.1126/science.281.5378.825.

Machado MP, Matos I, Grosso AR, Schartl M, Coelho MM. 2016. Non-canonical expression patterns and evolutionary rates of sex-biased genes in a seasonal fish. Mol. Reprod. Dev. 83:1102–1115. doi: 10.1002/mrd.22752.

Marshall OJ. 2004. PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. Bioinformatics. 20:2471–2472. doi: 10.1093/bioinformatics/bth254.

Matos IMN, Coelho MM, Schartl M. 2016. Gene copy silencing and DNA methylation in natural and artificially produced allopolyploid fish. J. Exp. Biol. 219:jeb.140418. doi: 10.1242/jeb.140418.

McClung CR. 2011. Circadian Rhythms: Lost in Post-Translation. Curr. Biol. 21:R400–R402. doi: 10.1016/j.cub.2011.04.024.

von Mering C et al. 2003. STRING: A database of predicted functional associations between proteins. Nucleic Acids Res. 31:258–261. doi: 10.1093/nar/gkg034.

Murrell B et al. 2012. Detecting individual sites subject to episodic diversifying selection Malik, HS, editor. PLoS Genet. 8:e1002764. doi: 10.1371/journal.pgen.1002764.

Murrell B et al. 2015. Gene-wide identification of episodic selection. Mol. Biol. Evol. 32:1365–1371. doi: 10.1093/molbev/msv035.

Muschick M, Indermaur A, Salzburger W. 2012. Convergent Evolution within an Adaptive Radiation of Cichlid Fishes. Curr. Biol. 22:2362–2368. doi: 10.1016/j.cub.2012.10.048.

Nakahata Y et al. 2008. A direct repeat of E-box-like elements is required for cell-autonomous circadian rhythm of clock genes. BMC Mol. Biol. 9:1. doi: 10.1186/1471-2199-9-1.

Nath A, Chaube R, Subbiah K. 2013. An insight into the molecular basis for convergent evolution in fish antifreeze Proteins. Comput. Biol. Med. 43:817–821. doi: 10.1016/j.compbiomed.2013.04.013.

Oates ME et al. 2015. The SUPERFAMILY 1.75 database in 2014: a doubling of data. Nucleic Acids Res. 43:D227–D233. doi: 10.1093/nar/gku1041.

Pando MP, Pinchak AB, Cermakian N, Sassone-Corsi P. 2001. A cell-based system that recapitulates the dynamic light-dependent regulation of the vertebrate clock. Proc. Natl. Acad. Sci. U. S. A. 98:10178–10183. doi: 10.1073/pnas.181228598.

Pando MP, Sassone-Corsi P. 2002. Unraveling the mechanisms of the vertebrate circadian clock: zebrafish may light the way. BioEssays. 24:419–426. doi: 10.1002/bies.10091.

Panja AS, Bandopadhyay B, Maiti S. 2015. Protein Thermostability Is Owing to Their Preferences to Non-Polar Smaller Volume Amino Acids, Variations in Residual Physico-Chemical Properties and More Salt-Bridges. PLoS ONE. 10. doi: 10.1371/journal.pone.0131495.

Paranjpe DA, Sharma VK. 2005. Evolution of temporal order in living organisms. J. Circadian Rhythms. 3:7. doi: 10.1186/1740-3391-3-7.

Passow CN, Arias-Rodriguez L, Tobler M. 2017. Convergent evolution of reduced energy demands in extremophile fish. PLoS ONE. 12. doi: 10.1371/journal.pone.0186935.

Ponting CP, Aravind L. 1997. PAS: a multifunctional domain family comes to light. Curr. Biol. CB. 7:R674–7. doi: 10.1016/S0960-9822(06)00352-6.

Potter SC et al. 2018. HMMER web server: 2018 update. Nucleic Acids Res. 46:W200–W204. doi: 10.1093/nar/gky448.

Prakash A, Jeffryes M, Bateman A, Finn RD. 2017. The HMMER Web Server for Protein Sequence Similarity Search. Curr. Protoc. Bioinforma. 60:3.15.1-3.15.23. doi: 10.1002/cpbi.40.

Preitner N et al. 2002. The Orphan Nuclear Receptor REV-ERBα Controls Circadian Transcription within the Positive Limb of the Mammalian Circadian Oscillator. Cell. 110:251–260. doi: 10.1016/S0092-8674(02)00825-5.

Protas ME et al. 2006. Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. Nat. Genet. 38:107–111. doi: 10.1038/ng1700.

Reischl S, Kramer A. 2011. Kinases and phosphatases in the mammalian circadian clock. FEBS Lett. 585:1393–1399. doi: 10.1016/j.febslet.2011.02.038.

Reppert SM, Weaver DR. 2001. Molecular Analysis of Mammalian Circadian Rhythms. Annu. Rev. Physiol. 63:647–676. doi: 10.1146/annurev.physiol.63.1.647.

Ronquist F et al. 2012. Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542. doi: 10.1093/sysbio/sys029.

Selby CP, Sancar A. 2006. A cryptochrome/photolyase class of enzymes with single-stranded DNA-specific photolyase activity. Proc. Natl. Acad. Sci. U. S. A. 103:17696–17700. doi: 10.1073/pnas.0607993103.

Shental-Bechor D, Levy Y. 2008. Effect of glycosylation on protein folding: A close look at thermodynamic stabilization. Proc. Natl. Acad. Sci. U. S. A. 105:8256–8261. doi: 10.1073/pnas.0801340105.

Sigrist CJA et al. 2013. New and continuing developments at PROSITE. Nucleic Acids Res. 41. doi: 10.1093/nar/gks1067.

Sigrist CJA et al. 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief. Bioinform. 3:265–274. doi: 10.1093/bib/3.3.265.

Smith MD et al. 2015. Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol. Biol. Evol. 32:1342–1353. doi: 10.1093/molbev/msv022.

Söti C, Pál C, Papp B, Csermely P. 2005. Molecular chaperones as regulatory elements of cellular networks. Curr. Opin. Cell Biol. 17:210–215. doi: 10.1016/j.ceb.2005.02.012.

Szklarczyk D et al. 2015. STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 43:D447–D452. doi: 10.1093/nar/gku1003.

Tamai TK, Young LC, Whitmore D. 2007. Light signaling to the zebrafish circadian clock by Cryptochrome 1a. Proc. Natl. Acad. Sci. U. S. A. 104:14712–14717. doi: 10.1073/pnas.0704588104.

Tamaru T et al. 2011. Synchronization of Circadian Per2 Rhythms and HSF1-BMAL1:CLOCK Interaction in Mouse Fibroblasts after Short-Term Heat Shock Pulse. PLOS ONE. 6:e24521. doi: 10.1371/journal.pone.0024521.

The UniProt Consortium. 2015. UniProt: a hub for protein information. Nucleic Acids Res. 43:D204–D212. doi: 10.1093/nar/gku989.

Toloza-Villalobos J, Arroyo JI, Opazo JC. 2015. The Circadian Clock of Teleost Fish: A Comparative Analysis Reveals Distinct Fates for Duplicated Genes. J. Mol. Evol. 80:57–64. doi: 10.1007/s00239-014-9660-x.

Trimmel H et al. 2018. Can riparian vegetation shade mitigate the expected rise in stream temperatures due to climate change during heat waves in a human-impacted pre-alpine river? Hydrol. Earth Syst. Sci. 22:437–461. doi: 10.5194/hess-22-437-2018.

Tsuchiya Y, Akashi M, Nishida E. 2003. Temperature compensation and temperature resetting of circadian rhythms in mammalian cultured fibroblasts. Genes Cells. 8:713–720. doi: 10.1046/j.1365-2443.2003.00669.x.

Vallone D, Gondi SB, Whitmore D, Foulkes NS. 2004. E-box function in a period gene repressed by light. Proc. Natl. Acad. Sci. 101:4106–4111. doi: 10.1073/pnas.0305436101.

Vatine G et al. 2009. Light Directs Zebrafish period2 Expression via Conserved D and E Boxes Kramer, A, editor. PLoS Biol. 7:e1000223. doi: 10.1371/journal.pbio.1000223.

Vaze KM, Sharma VK. 2013. On the adaptive significance of circadian clocks for their owners. Chronobiol. Int. 30:413–33. doi: 10.3109/07420528.2012.754457.

Waap S, Amaral AR, Gomes B, Manuela Coelho M. 2011. Multi-locus species tree of the chub genus Squalius (Leuciscinae: Cyprinidae) from western Iberia: new insights into its evolutionary history. Genetica. 139:1009–1018. doi: 10.1007/s10709-011-9602-0.

Wallace IM, O'Sullivan O, Higgins DG, Notredame C. 2006. M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res. 34:1692–1699. doi: 10.1093/nar/gkl091.

Wang H. 2008a. Comparative analysis of period genes in teleost fish genomes. J. Mol. Evol. 67:29–40. doi: 10.1007/s00239-008-9121-5.

Wang H. 2008b. Comparative analysis of teleost fish genomes reveals preservation of different ancient clock duplicates in different fishes. Mar. Genomics. 1:69–78. doi: 10.1016/j.margen.2008.06.003.

Wang H. 2009. Comparative genomic analysis of teleost fish bmal genes. Genetica. 136:149–161. doi: 10.1007/s10709-008-9328-9.

Wang M, Zhong Z, Zhong Y, Zhang W, Wang H. 2015. The Zebrafish Period2 Protein Positively Regulates the Circadian Clock through Mediation of Retinoic Acid Receptor (RAR)-related Orphan Receptor α (Rorα). J. Biol. Chem. 290:4367–4382. doi: 10.1074/jbc.M114.605022.

Weaver S et al. 2018. Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. Mol. Biol. Evol. 35:773–777. doi: 10.1093/molbev/msx335.

Weger BD et al. 2011. The Light Responsive Transcriptome of the Zebrafish: Function and Regulation Mueller, F, editor. PLoS ONE. 6:e17080. doi: 10.1371/journal.pone.0017080.

Whitmore D, Foulkes NS, Strähle U, Sassone-Corsi P. 1998. Zebrafish Clock rhythmic expression reveals independent peripheral circadian oscillators. Nat. Neurosci. 1:701–707. doi: 10.1038/3703.

Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. Springer International Publishing //www.springer.com/la/book/9783319242750 (Accessed November 26, 2018).

Wu CH. 2004. PIRSF: family classification system at the Protein Information Resource. Nucleic Acids Res. 32:112D – 114. doi: 10.1093/nar/gkh097.

Yoshizawa-Sugata N, Masai H. 2007. Human Tim/Timeless-interacting protein, Tipin, is required for efficient progression of S phase and DNA replication checkpoint. J. Biol. Chem. 282:2729–2740. doi: 10.1074/jbc.M605596200.

Zhang W et al. 1995. Molecular cloning and characterization of NF-IL3A, a transcriptional activator of the human interleukin-3 promoter. Mol. Cell. Biol. 15:6055–6063. doi: 10.1128/MCB.15.11.6055.

Supplementary Material

Tables

Table S1. Environmental conditions observed for each basin between 2001 and 2016. Temperature and pH data were obtained from snirh.pt (National Information System of Water Resources) and number of daylight hours was obtained from Time and Date AS (timeanddate.com; accessed in June 2018).

Basin	Average water temperature (ºC)				Average water pH	Average light hours per day			
	Winter	Spring	Summer	Autumn		Winter	Spring	Summer	Autumn
Mondego	8.8	16.0	20.9	14.1	7.0	10:42	13:30	13:36	10:48
Tagus	11.5	15.1	19.5	16.1	7.3	10:42	13:30	13:36	10:48
Mira	15.9	17.8	24.6	23.2	7.7	10:54	13:45	13:52	10:54
Arade	13.8	20.9	24.3	19.1	8.7	10:54	13:47	13:52	10:54
Almargem	13.0	19.2	*NA*	18.9	8.1	10:54	13:47	13:52	10:54

Table S2. List of *Danio rerio* Uniprot accession ID for target proteins and ENA accession IDs for corresponding coding genes.

Protein	Uniprot accession ID	ENA accession ID of coding gene
CRY1AA	A4QN37	AAI63354
CRY1AB	Q9I916	AB042249
CRY1BA	B0S7X5	BC095305
CRY1BB	Q4V9Q3	BC164795
CRY2	Q9I913	AB042252
CRY3	Q9I912	AB042253
PER1A	Q7SZZ4	BC163543
PER1B	B3DK47	AAI63718
PER2	Q7T1C9	AAO38747
PER3	Q9I8L4	AAF87986
CLOCKA	Q9W6J4	AAD27749
CLOCKB	B3DH92	AAI62682
BMAL1A	Q9I879	AAF64395
BMAL1B	Q8JIG1	BAC02688
BMAL2	Q9IAU1	AAF64396
TIMELESS	E7FGL0	

Table S3. List of primer pairs used in PCR to re-sequence circadian-related genes with Sanger method in *Squalius* species.

Genes	Primer name (1)	Primer Sequence (1)	Primer name (2)	Primer Sequence (2)
cry1aa	cry1_f1	Forward: 5' – GCGGACTGGAGTGTGAA – 3'	cry1_fw2	Forward: 5' – AGATCTACCAGCAACTGT – 3'
	cry1_r1	Reverse: 5' – TGTGCAGATTACAGAGCC – 3'	cry1_rv2	Reverse: 5' – AGAGCTGCTCTGACGTTGG – 3'
cry1ab	cry1ab_fw1	Forward: 5' – ATGGTTGTCAATACGATC – 3'	cry1ab_fw2	Forward: 5' – GTGGGACAGGAATCCC – 3'
	cry1ab_rv1	Reverse: 5' – CCACTTCGCCAAGGCTTC – 3'	cry1ab_rv2	Reverse: 5' – GTAACTATATTGCCTCT – 3'
cry1ba	cry2a_fw1	Forward: 5' – TCCACTCTCCCTCTACGGT – 3'	cry2a_fw2	Forward: 5' – GGACCAATACACCAGCAC – 3'
	cry2a_rv1	Reverse: 5' – ACTCCTGTTACCACTGGC – 3'	cry2a_rv2	Reverse: 5' – GCATCGCAGCAGAGTATG – 3'
cry1bb	Cry1bb_fw1	Forward: 5' – ATGTCCTCCATCAACTCGG – 3'	Cry1bb_fw2	Forward: 5' – ATCTCTCTGTACGACAAGC – 3'
	Cry1bb_rv1	Reverse: 5' – GGTTATTAGTGGCGGCCGT – 3'	Cry1bb_rv2	Reverse: 5' – CTGCATCTTCCTGCCTCTT– 3'
cry2	cry3_fw1	Forward: 5' – CAAACTTCACAGAGAGGCG – 3'	cry3_fw2	Forward: 5' – CCTACGATTCGGCTGTCT – 3'
	cry3_rv1	Reverse: 5' – CACGACATGACAGACAGC – 3'	cry3_rv2	Reverse: 5' – TGAAGTCTGAACCTCTGGC – 3'
cry3	cry4_fw1	Forward: 5' – CGTTCGGTATGTCACATG – 3'	cry4_fw2	Forward: 5' – GCTGATCCACAGGTCTCCT – 3'
	cry4_rv1	Reverse: 5' – TGAAGGTGTTCGAGGAGT – 3'	cry4_rv2	Reverse: 5' – TCTCCTCCGCTCACGTACA – 3'
per1a	Per1a_f1	Forward: 5' – CGACAATCAATCTCTATG – 3'	Per1a_fw2	Forward: 5' – ATAAGAAAATTCTTCAGTT – 3'
	Per1a_r1	Reverse: 5' – GAGCAGTCTGTCTTCTGGAT – 3'	Per1a_rv2	Reverse: 5' – GATTCTGGTCATTCTCTGA – 3'
per1b	Per1b_fw1	Forward: 5' – ATGAGCGATGATAACTCCG – 3'	Per1b_fw2	Forward: 5' – CATACCCATGGGACCTGTAG – 3'
	Per1b_rv1	Reverse: 5' – GAATTAGAGGGCCTGTCCA – 3'	Per1b_rv2	Reverse: 5' – TGCATGACTCTTCTTCTC – 3'
per2	Per2_fw1	Forward: 5' – TGTCTGAGGACTCAGATTCC – 3'	Per2_fw2	Forward: 5' – ATAAGAAAATTCTTCAGTT – 3'
	Per2_rv1	Reverse: 5' – CGAGACCCTGACCTTCAGTG – 3'	Per2_rv2	Reverse: 5' – GATTCTGGTCATTCTCTGA – 3'
per3	per3_fw1	Forward: 5' – TTCCCGATGGAGAAGAGGA – 3'	per3_fw2	Forward: 5' – GACTGGAGTTAGGATGAAT – 3'
	per3_rv1	Reverse: 5' – GACTGGAGTTAGGATGAATG – 3'	per3_rv2	Reverse: 5' – TCATGGGCTGGAGGTTCTG – 3'
clocka	clock1_fw1	Forward: 5' – TGACCTCCAGCATAGACC – 3'	clock1_fw2	Forward: 5' – CACCTCCAGACACAGTTTAG – 3'
	clock1_rv1	Reverse: 5' – CCTGAAGTACCCAGAACCTC – 3'	clock1_rv2	Reverse: 5' – GGATGTTAGCCTCAATCATG – 3'
clockb	clockb_fw1	Forward: 5' – ATGAGTTCCAGAGACAGG – 3'	clockb_fw2	Forward: 5' – TCTCGACACAGTCTGGAG – 3'
	clockb_rv1	Reverse: 5' – GGTGGTGCTCTGTGATCT – 3'	clockb_rv2	Reverse: 5' – GTTGAACCTGCTGCTGTGG – 3'
bmal1a	bmal1a_fw1	Forward: 5' – TCCTCCACGATGAATGAG – 3'	bmal1a_fw2	Forward: 5' – ACAGCACTGGCTATCTGA – 3'
	bmal1a_rv1	Reverse: 5' – GACGATATGCGGATGAAG – 3'	bmal1a_rv2	Reverse: 5' – GGAGGCTCATGATAACAG – 3'
bmal1b	bmal1b_fw1	Forward: 5' – TGGCAGACCAAAGAATGGA – 3'	bmal1b_fw2	Forward: 5' – GCTATCTGAAGAGCTGGC – 3'
	bmal1b_rv1	Reverse: 5' – TTCATCTAAGCCCATCTTAA – 3'	bmal1b_rv2	Reverse: 5' – CAGGGTAAGTCGCTGAAGT – 3'
bmal2	Bmal2_f1	Forward: 5' – GTGGTCGAGGAACACAGCG – 3'	Bmal2_f2	Forward: 5' – GAGTCCCAGCGGTACTGCA – 3'
	Bmal2_rv1	Reverse: 5' – GCACGGTGCAGTACCGCTG – 3'	Bmal2_rv2	Reverse: 5' – CAGACCAGTGCATCTCGTC – 3'
timeless	tim_fw1	Forward: 5' – GGACTTGTACATGATGAACT – 3'	tim_fw2	Forward: 5' – TCTGAAGCAGATTCTCTTCA – 3'
	tim_rv1	Reverse: 5' – GCTTCAGAAGTTCTAGTTCC – 3'	tim_rv2	Reverse: 5' – GCTTCTGTCAAACTTACTCC – 3'

Table S4. PCR conditions for each pair of primers used in amplification of circadian-related genes.

Genes	Pair of primers	Initial Denaturation		PCR Cycles							Final Extension	
				Denaturation		Annealing		Extension				
		Temp (ºC)	Time (s)	Temp (ºC)	Time (s)	Temp (ºC)	Time (s)	Temp (ºC)	Time (s)	Cycles	Temp (ºC)	Time (s)
cry1aa	(1)	95	300	95	45	55	60	72	45	30	72	600
cry1ab	(1)	95	300	95	45	54	60	72	60	32	72	600
cry1ba	(1)	95	300	95	45	55	60	72	45	30	72	600
cry1bb	(1)	95	300	95	45	55	60	72	45	35	72	600
cry2	(1)/(2)	95	300	95	45	56	60	72	60	30	72	600
cry3	(1)	95	300	95	45	54	60	72	45	32	72	600
cry1aa	(2)	95	300	95	45	56	60	72	45	30	72	600
cry1ab	(2)	95	300	95	45	56	60	72	60	32	72	600
cry1ba	(2)	95	300	95	45	54	60	72	45	30	72	600
cry1bb	(2)	95	300	95	45	57	60	72	45	35	72	600
cry3	(2)	95	300	95	45	58	60	72	45	32	72	600
per1a	(1)/(2)	95	300	95	45	56	60	72	45	30	72	600
per1b	(1)/(2)	95	300	95	45	56	60	72	60	32	72	600
per2	(1)/(2)	95	300	95	45	57	60	72	45	30	72	600
per3	(1)/(2)	95	300	95	45	54	60	72	45	35	72	600
clocka	(1)/(2)	95	300	95	45	60	60	72	60	30	72	600
clockb	(1)/(2)	95	300	95	45	54	60	72	45	32	72	600
bmal1a	(1)/(2)	95	300	95	45	56	60	72	60	32	72	600
bmal1b	(1)/(2)	95	300	95	45	56	60	72	60	32	72	600
bmal2	(1)	95	300	95	45	55	60	72	60	35	72	600
bmal2	(2)	95	300	95	45	58	60	72	45	30	72	600
timeless	(1)	95	300	95	45	53	60	72	60	35	72	600
timeless	(2)	95	300	95	45	54	60	72	60	34	72	600

Table S5. List of *Drosophila melanogaster* Uniprot accession ID for protein sequences used as outgroup in phylogenetic analysis and ENA accession IDs for corresponding coding genes.

Protein	Uniprot accession ID	ENA accession ID of coding gene
CRYPTOCHROME	O77059	AF099734
PERIOD	P07663	AAB87476
CLOCK	O61735	AF065133
CYCLE	O61734	AF069998

Table S6. Description and biological relevance of physicochemical parameters of proteins analysed

Physicochemical parameter	**Description and biological relevance**	**References**
Theoretical Isoelectric point (pI)	Isoelectric point of a protein is the pH at which a protein has no net charge. pI and net charge of a protein is important for its solubility, subcellular localization, and possible interactions.	(Gasteiger *et al.* 2005; Kiraga *et al.* 2007; Khaldi & Shields 2011)
Instability index	Estimate of the stability of your protein in a test tube. A protein whose instability index is smaller than 40 is considered as stable; a value above 40 predicts that the protein may be unstable.	(Gasteiger *et al.* 2005)
Aliphatic index	The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains (Ala, Val, Ile, Leu). It may be regarded as a positive factor for the increase of thermostability of globular proteins.	(Ikai 1980; Gasteiger *et al.* 2005)

Table S7. Summary of the results obtained by MEME analysis for episodic positive selection. A threshold of 0.1 was assumed for significance level. The type of mutation was inferred by analysing the physicochemical properties of the amino acids substituted and the ancestral form was inferred from zebrafish amino acid at that position.

Protein	Number of sites under positive selection	Sites under positive selection	p-value	Type of non-synonymous mutation
CRY1AA	1	K275Q	0.03	Non-conservative
CRY1AB	0	-	-	-
CRY1BA	0	-	-	-
CRY1BB	0	-	-	-
CRY2	0	-	-	-
CRY3	0	-	-	-
PER1A	1	T254A	0.08	Non-conservative
PER1B	1	T1256V	0.08	Non-conservative
PER2	4	T402D	0.01	Non-conservative
		T514L	0.02	Non-conservative
		R697A	0.03	Non-conservative
		G1208T	0.09	Non-conservative
PER3	1	K429Y	0.02	Non-conservative
CLOCKA	0	-	-	-
CLOCKB	0	-	-	-
BMAL1A	0	-	-	-
BMAL1B	0	-	-	-
BMAL2	1	R20P	0.09	Non-conservative
TIMELESS	2	L360N	0.01	Non-conservative
		F682Y	0.03	Conservative

Table S8. Summary of the results obtained by FEL analysis for pervasive negative selection in coding genes for CRY proteins. A threshold of 0.1 was assumed for significance level.

Protein	Number of sites under negative selection	Sites under negative selection	p-value
CRY1AA	16	316V	0.092
		348R	0.069
		349Q	0.056
		355H	0.034
		360A	0.090
		361V	0.045
		366T	0.075
		369K	0.054
		380V	0.095
		381F	0.034
		384L	0.080
		387D	0.031
		388A	0.055
		393N	0.032
		415P	0.079
		565L	0.056
CRY1AB	2	191G	0.097
		597H	0.034

Table S8. (continued)

Protein	Number of sites under negative selection	Sites under negative selection	p-value
CRY1BA	3	422T	0.089
		427D	0.047
		661P	0.099
CRY1BB	12	61E	0.011
		82A	0.091
		203T	0.086
		209A	0.082
		294L	0.089
		312R	0.089
		345D	0.008
		347I	0.078
		365V	0.077
		475V	0.055
		509S	0.080
		595G	0.080
CRY2	24	128N	0.007
		203R	0.100
		215T	0.098
		347L	0.090
		348R	0.036
		350E	0.068
		358R	0.092
		360A	0.091
		361V	0.089
		364F	0.049
		373S	0.039
		375E	0.009
		379K	0.005
		389D	0.050
		402C	0.046
		407Q	0.081
		412C	0.052
		415P	0.099
		418F	0.017
		424P	0.080
		437E	0.075
		438Y	0.044
		463G	0.062
		567R	0.015

Table S8. (continued)

Protein	Number of sites under negative selection	Sites under negative selection	p-value
CRY3	18	67L	0.008
		75Y	0.020
		81T	0.001
		97I	0.037
		156H	0.049
		221H	0.029
		262F	0.002
		338D	0.030
		345R	0.073
		346Q	0.034
		356H	0.029
		357A	0.081
		358V	0.058
		384D	0.021
		397L	0.034
		399A	0.092
		400S	0.018
		429Y	0.030

Table S9. Analysis of patters of post-translational modifications for CRY proteins. Rows shaded in grey correspond to PTMs with modifications in studied populations.

		Species				
Protein	**Post-translational modification/Domain**	*S. torgalensis*	*S. aradensis*	*S. carolitertii*	*S. pyrenaicus (Tagus)*	*S. pyrenaicus (Almargem)*
CRY1AA	*Protein kinase C phosphorylation site*	5	5	5	5	5
	Casein kinase II phosphorylation site	9	9	9	9	9
	N-glycosylation site	4	4	4	4	4
	cAMP- and cGMP-dependent protein kinase phosphorylation site	2	2	2	2	2
	Cell attachment sequence	1	1	1	1	1
	Amidation site	2	2	2	2	2
CRY1AB	*Protein kinase C phosphorylation site*	5	5	5	5	6
	Casein kinase II phosphorylation site	10	10	11	11	10
	cAMP- and cGMP-dependent protein kinase phosphorylation site	2	2	2	2	2
	Cell attachment sequence	1	1	1	1	1
	Amidation site	1	1	1	1	1
CRY1BA	*Protein kinase C phosphorylation site*	7	7	6	6	6
	Casein kinase II phosphorylation site	7	7	7	7	7
	cAMP- and cGMP-dependent protein kinase phosphorylation site	3	3	2	2	2
	Cell attachment sequence	1	1	1	1	1
	Amidation site	1	1	2	2	2
	N-glycosylation site	3	3	3	3	3
	ATP/GTP-binding site motif A (P-loop)	1	1	-	-	-

Table S9. (continued)

Protein	Post-translational modification/Domain	Species				
		S. torgalensis	*S. aradensis*	*S. carolitertii*	*S. pyrenaicus (Tagus)*	*S. pyrenaicus (Almargem)*
CRY1BB	*Protein kinase C phosphorylation site*	9	7	10	9	8
	Casein kinase II phosphorylation site	8	8	8	8	8
	Tyrosine kinase phosphorylation site	1	1	1	1	1
	Cell attachment sequence	2	2	2	2	2
	Amidation site	1	2	2	2	2
	N-glycosylation site	1	-	1	1	-
CRY2	*Protein kinase C phosphorylation site*	7	7	7	7	7
	Casein kinase II phosphorylation site	6	6	6	6	6
	Cell attachment sequence	2	2	2	2	2
	Amidation site	1	1	1	1	1
	N-glycosylation site	1	1	1	1	1
	cAMP- and cGMP-dependent protein kinase phosphorylation site	2	2	2	2	2
CRY3	*Protein kinase C phosphorylation site*	8	6	8	8	7
	Casein kinase II phosphorylation site	10	10	10	10	10
	Cell attachment sequence	1	1	1	1	1
	Amidation site	1	1	1	1	1
	N-glycosylation site	2	2	2	2	2
	cAMP- and cGMP-dependent protein kinase phosphorylation site	-	-	1	1	1

Table S10. Patterns of protein-protein interactions for CRY protein predicted with STRING with a threshold of 0.7 for score. Rows shaded in grey correspond to circadian-related proteins studied here; in orange are highlighted proteins related to temperature responses; in blue are highlighted putative circadian proteins with secondary functions

Protein	Interactors	Interactor description	Score
CRY1AA	PER2	Period circadian clock 2	0.998
	BMAL1A	aryl hydrocarbon receptor nuclear translocator-like 1a	0.988
	PER1B	Period circadian clock 1B	0.982
	PER3	Period circadian clock 3	0.977
	PER1A	Period circadian clock 1A	0.957
	CLOCKB	Clock circadian clock regulator b	0,930
	TIMELESS	Timeless circadian clock	0.927
	BHLHe41	basic helix-loop-helix family, member e41	0.914
	TEFa	Thyrotrophic embryonic factor alpha	0.904
	BMAL1B	aryl hydrocarbon receptor nuclear translocator-like 1b	0.898
CRY1AB	PER2	Period circadian clock 2	0.968
	PER1B	Period circadian clock 1B	0.964
	PER1A	Period circadian clock 1A	0.950
	PER3	Period circadian clock 3	0.944
	BMAL1A	aryl hydrocarbon receptor nuclear translocator-like 1a	0.923
	BMAL1B	aryl hydrocarbon receptor nuclear translocator-like 1b	0.906
	TIMELESS	Timeless circadian clock	0,850
	CLOCKB	Clock circadian clock regulator b	0.849
	CLOCKA	Clock circadian clock regulator a	0.798
	BHLHe41	basic helix-loop-helix family, member e41	0.793

Table S10. (continued)

Protein	Interactors	Interactor description	Score
CRY1BA	BMAL1A	aryl hydrocarbon receptor nuclear translocator-like 1a	0.984
	PER2	Period circadian clock 2	0.982
	PER1B	Period circadian clock 1B	0.978
	PER3	Period circadian clock 3	0.977
	PER1A	Period circadian clock 1A	0.956
	BMAL1B	aryl hydrocarbon receptor nuclear translocator-like 1b	0.939
	CLOCKB	Clock circadian clock regulator b	0.910
	BHLHe41	basic helix-loop-helix family, member e41	0.904
	CLOCKA	Clock circadian clock regulator a	0.882
	CSNK1E	Casein kinase I isoform epsilon	0.827
CRY1BB	PER3	Period circadian clock 3	0.944
	PER2	Period circadian clock 2	0.928
	PER1B	Period circadian clock 1B	0.910
	BMAL1A	aryl hydrocarbon receptor nuclear translocator-like 1a	0.900
	PER1A	Period circadian clock 1A	0.885
	BMAL1B	aryl hydrocarbon receptor nuclear translocator-like 1b	0.869
	CLOCKB	Clock circadian clock regulator b	0.819
	CLOCKA	Clock circadian clock regulator a	0.812
	NR1D2B	Nuclear receptor subfamily 1, group D, member 2B	0.703
	TIMELESS	Timeless circadian clock	0.700
CRY2	NR1D1	Nuclear receptor subfamily 1, group D, member 1	0.985
	HSF2	Heat-shock transcription factor 2	0.979
	CIART	Circadian-associated repressor of transcription a	0.977
	PER2	Period circadian clock 2	0.976
	PER3	Period circadian clock 3	0.969
	ZNF395B	Zinc finger protein 395b	0.968
	BHLHe41	basic helix-loop-helix family, member e41	0.958
	CLOCKB	Clock circadian clock regulator b	0.955
	TIMELESS	Timeless circadian clock	0.952
	BMAL1A	aryl hydrocarbon receptor nuclear translocator-like 1a	0.948
CRY3	BMAL1A	aryl hydrocarbon receptor nuclear translocator-like 1a	0.993
	PER2	Period circadian clock 2	0.923
	PER1B	Period circadian clock 1B	0.921
	PER3	Period circadian clock 3	0.908
	CLOCKB	Clock circadian clock regulator b	0.898
	TIMELESS	Timeless circadian clock	0.894
	PER1A	Period circadian clock 1A	0.846
	BMAL1B	aryl hydrocarbon receptor nuclear translocator-like 1b	0.835
	CLOCKA	Clock circadian clock regulator a	0.759
	NFIL3-5	nuclear factor, interleukin 3-regulates, member 5	0.750

Table S11. Summary of the results obtained by FEL analysis for pervasive negative selection in coding genes of PER proteins. A threshold of 0.1 was assumed for significance level.

Protein	Number of sites under negative selection	Sites under negative selection	p-value
PER1A	5	E16	0.047
		K670	0.053
		K1128	0.039
		L1190	0.076
		P1216	0.070
PER1B	17	V341	0.004
		P351	0.099
		P509	0.100
		T595	0.061
		S641	0.026
		L671	0.079
		R903	0.090
		S939	0.057
		I940	0.061
		Y945	0.067
		G948	0.071
		L952	0.067
		M992	0.062
		A1042	0.068
		S1185	0.034
		D1252	0.090
		Y1259	0.042
PER2	59	9P	0.099
		34G	0.086
		56P	0.021
		78S	0.002
		79I	0.064
		103S	0.010
		241I	0.058
		244I	0.056
		247I	0.058
		368E	0.035
		391R	0.089
		405P	0.002
		457Q	0.023
		458P	0.087
		468R	0.078
		471E	0.039
		472Y	0.044
		474T	0.093
		476D	0.033
		477T	0.010
		483V	0.011
		485P	0.010
		489K	0.006
		492F	0.003

Table S11. (continued)

Protein	Number of sites under negative selection	Sites under negative selection	p-value
		512P	0.086
		513T	0.020
		520I	0.053
		521D	0.033
		523D	0.033
		524I	0.058
		525Q	0.072
		546G	0.092
		548G	0.078
		561S	0.096
		565S	0.032
		566N	0.017
		568N	0.017
		569G	0.093
		595H	0.027
		647K	0.065
		654Y	0.070
PER2	59	656Q	0.044
		657I	0.081
		658S	0.049
		659C	0.035
		670L	0.029
		681Q	0.072
		686T	0.094
		748S	0.029
		752I	0.081
		824F	0.027
		925Q	0.038
		949G	0.088
		1011F	0.034
		1210N	0.026
		1258A	0.087
		1298K	0.032
		1370E	0.043
		1394C	0.034

Table S11. (continued)

Protein	Number of sites under negative selection	Sites under negative selection	p-value
PER3	19	29T	0.096
		33P	0.011
		70H	0.031
		77S	0.036
		162V	0.086
		308F	0.033
		381T	0.092
		425T	0.091
		503C	0.037
		579Q	0.029
		613I	0.072
		627H	0.041
		768T	0.092
		778T	0.094
		929Q	0.039
		1105S	0.033
		1140S	0.012
		1212T	0.095
		1257D	0.041

Table S12. Analysis of patters of post-translational modifications for PER proteins. Rows shaded in grey correspond to PTMs with modifications in studied populations.

		Species				
Protein	**Post-translational modification/Domain**	*S. torgalensis*	*S. aradensis*	*S. carolitertii*	*S. pyrenaicus (Tagus)*	*S. pyrenaicus (Almargem)*
PER1A	*Protein kinase C phosphorylation site*	12	11	12	12	11
	Casein kinase II phosphorylation site	25	25	25	25	25
	N-glycosylation site	9	10	12	12	11
	cAMP- and cGMP-dependent protein kinase phosphorylation site	2	2	1	1	1
	Amidation site	1	1	1	1	1
PER1B	*Protein kinase C phosphorylation site*	19	19	19	19	19
	Casein kinase II phosphorylation site	27	26	27	27	27
	N-glycosylation site	14	14	14	14	14
	cAMP- and cGMP-dependent protein kinase phosphorylation site	1	1	1	1	1
	Amidation site	3	3	3	3	3
	Tyrosine kinase phosphorylation site	1	1	1	1	1

Table S12. (continued)

		Species				
Protein	**Post-translational modification/Domain**	*S. torgalensis*	*S. aradensis*	*S. carolitertii*	*S. pyrenaicus (Tagus)*	*S. pyrenaicus (Almargem)*
PER2	*Protein kinase C phosphorylation site*	24	25	24	24	23
	Casein kinase II phosphorylation site	38	37	38	38	36
	N-glycosylation site	15	15	15	15	15
	cAMP- and cGMP-dependent protein kinase phosphorylation site	1	2	2	2	2
	Amidation site	2	1	2	2	2
	Tyrosine kinase phosphorylation site	1	-	1	1	-
PER3	*Protein kinase C phosphorylation site*	18	19	17	17	18
	Casein kinase II phosphorylation site	24	24	24	24	23
	N-glycosylation site	10	10	10	11	11
	Tyrosine kinase phosphorylation site	1	1	1	1	1
	cAMP- and cGMP-dependent protein kinase phosphorylation site	2	2	2	2	2
	Cell attachment sequence	1	1	1	1	1

Table S13. Patterns of protein-protein interactions for PER proteins predicted with STRING with a threshold of 0.7 for score. Rows shaded in grey correspond to circadian-related proteins studied here; in orange are highlighted protein related to temperature responses; in blue are highlighted putative circadian proteins with secondary functions

Protein	Interactors	Interactor description	Score
PER1A	BHLHe41	basic helix-loop-helix family, member e41	0.959
	CRY1AA	Cryptochome circadian clock 1AA	0.957
	CRY1BA	Cryptochome circadian clock 1BA	0.956
	CRY1BB	Cryptochome circadian clock 1BB	0,950
	PER3	Period circadian clock 3	0.922
	CSNK1E	Casein kinase I isoform epsilon	0.914
	CSNK1Db	casein kinase I, isoform delta b	0.914
	CRY2	Cryptochome circadian clock 2	0.907
	BMAL1A	aryl hydrocarbon receptor nuclear translocator-like 1a	0.905
PER1B	CRY1AA	Cryptochome circadian clock 1AA	0.982
	CRY1AB	Cryptochome circadian clock 1AB	0.978
	BMAL1A	aryl hydrocarbon receptor nuclear translocator-like 1a	0.975
	BHLHe41	basic helix-loop-helix family, member e41	0.966
	CRY1BB	Cryptochome circadian clock 1BB	0.964
	CRY5	Cryptochrome 5	0.952
	CRY2	Cryptochome circadian clock 2	0.949
	CSNK1Db	casein kinase I, isoform delta b	0.944
	CSNK1E	Casein kinase I isoform epsilon	0.943
	PER3	Period circadian clock 3	0.941

Table S13. (continued)

Protein	Interactors	Interactor description	Score
PER2	CRY1AA	Cryptochome circadian clock 1AA	0.998
	CRY5	Cryptochrome 5	0.990
	CRY2	Cryptochome circadian clock 2	0.982
	CRY1AB	Cryptochome circadian clock 1AB	0.976
	BMAL1A	aryl hydrocarbon receptor nuclear translocator-like 1a	0.976
	CRY1BB	Cryptochome circadian clock 1BB	0.968
	BHLHe41	basic helix-loop-helix family, member e41	0.959
	CRY-DASH	Cryptochrome DASH	0.949
	PER3	Period circadian clock 3	0.941
	CSNK1E	Casein kinase I isoform epsilon	0.937
PER3	BMAL1A	aryl hydrocarbon receptor nuclear translocator-like 1a	0.988
	CRY1AA	Cryptochome circadian clock 1AA	0.977
	CRY2	Cryptochome circadian clock 2	0.977
	CRY1BA	Cryptochome circadian clock 1BA	0.969
	CRY1BB	Cryptochome circadian clock 1BB	0.944
	CRY1AB	Cryptochome circadian clock 1AB	0.944
	CRY5	Cryptochrome 5	0.943
	PER1B	Period circadian clock 1B	0.941
	PER2	Period circadian clock 2	0.941
	PER1A	Period circadian clock 1A	0.922

Table S14. Summary of the results obtained by FEL analysis for pervasive negative selection in coding genes of CLOCK proteins. A threshold of 0.1 was assumed for significance level.

Protein	**Number of sites under negative selection**	**Sites under negative selection**	**p-value**
CLOCKA	15	V157	0.087
		I288	0.032
		P292	0.063
		T363	0.057
		T365	0.068
		E383	0.062
		S385	0.098
		A390	0.062
		Q395	0.075
		S397	0.100
		S399	0.100
		S401	0.100
		Q402	0.075
		A411	0.088
		Q485	0.075
CLOCKB	5	R29	0.087
		K168	0.056
		T194	0.041
		E254	0.038
		D305	0.019

Table S15. Analysis of patters of post-translational modifications for CLOCK proteins. Rows shaded in grey correspond to PTMs with modifications in studied populations.

		Species				
Protein	**Post-translational modification/Domain**	*S. torgalensis*	*S. aradensis*	*S. carolitertii*	*S. pyrenaicus (Tagus)*	*S. pyrenaicus (Almargem)*
CLOCKA	*Protein kinase C phosphorylation site*	13	14	13	13	14
	Casein kinase II phosphorylation site	15	15	15	15	15
	N-glycosylation site	4	4	4	4	4
	cAMP- and cGMP-dependent protein kinase phosphorylation site	3	3	3	3	3
CLOCKB	*Protein kinase C phosphorylation site*	10	10	10	10	10
	Casein kinase II phosphorylation site	10	10	10	10	10
	N-glycosylation site	5	5	5	5	5
	cAMP- and cGMP-dependent protein kinase phosphorylation site	3	3	3	3	3

Table S16. Patterns of protein-protein interactions for CLOCK proteins predicted with STRING with a threshold of 0.7 for score. Rows shaded in grey correspond to circadian-related proteins studied here.

Protein	Interactors	Interactor description	Score
CLOCKA	CLOCK2	Clock circadian clock regulator 2	0.990
	BMAL1A	aryl hydrocarbon receptor nuclear translocator-like 1a	0.953
	BMAL1B	aryl hydrocarbon receptor nuclear translocator-like 1b	0.952
	CRY1AA	Cryptochome circadian clock 1AA	0.933
	CRY1BA	Cryptochome circadian clock 1AB	0.930
	CRY2	Cryptochome circadian clock 2	0.930
	PER2	Period circadian clock 2	0.927
	PER1B	Period circadian clock 1B	0.925
	CRY5	Cryptochrome 5	0.925
CLOCKB	CLOCKA	Clock circadian clock regulator a	0.964
	CRY1BA	Cryptochome circadian clock 1BA	0.882
	CRY1AA	Cryptochome circadian clock 1AA	0.874
	PER2	Period circadian clock 2	0.850
	CRY1BB	Cryptochome circadian clock 1BB	0.816
	CRY1AB	Cryptochome circadian clock 1AB	0.812
	PER1B	Period circadian clock 1B	0.798
	PER3	Period circadian clock 3	0.795
	PER1A	Period circadian clock 1A	0.792

Table S17. Summary of the results obtained by FEL analysis for pervasive negative selection in coding genes of BMAL proteins. A threshold of 0.1 was assumed for significance level.

Protein	Number of sites under negative selection	Sites under negative selection	p-value
BMAL1A	4	T44	0.079
		K152	0.076
		A496	0.008
		P561	0.043
BMAL1B	4	A2	0.018
		G445	0.086
		G476	0.092
		S601	0.094
BMAL2	13	I116	0.089
		R201	0.089
		T241	0.080
		N244	0.011
		E353	0.038
		F354	0.041
		Y358	0.068
		F366	0.041
		Y377	0.068
		L399	0.064
		N460	0.059
		G503	0.083
		V580	0.079

Table S18. Analysis of patters of post-translational modifications for BMAL proteins. Rows shaded in grey correspond to PTMs with modifications in studied populations.

		Species				
Protein	**Post-translational modification/Domain**	*S. torgalensis*	*S. aradensis*	*S. carolitertii*	*S. pyrenaicus (Tagus)*	*S. pyrenaicus (Almargem)*
BMAL1A	*Protein kinase C phosphorylation site*	7	7	7	7	7
	Casein kinase II phosphorylation site	17	17	17	17	17
	cAMP- and cGMP-dependent protein kinase phosphorylation site	1	1	1	1	1
	N-glycosylation site	4	4	4	4	4
	Amidation site	2	2	2	2	2
BMAL1B	*Protein kinase C phosphorylation site*	7	7	7	7	7
	Casein kinase II phosphorylation site	15	15	15	15	15
	cAMP- and cGMP-dependent protein kinase phosphorylation site	2	2	2	2	2
	N-glycosylation site	5	5	5	5	5
	ATP/GTP-binding site motif A (P-loop)	1	1	1	1	1
	Amidation site	1	1	1	1	1
BMAL2	*Protein kinase C phosphorylation site*	9	9	8	8	8
	Casein kinase II phosphorylation site	17	17	18	18	18
	N-glycosylation site	1	1	1	1	1
	Tyrosine kinase phosphorylation site	2	2	2	2	2

Table S19. Patterns of protein-protein interactions for BMAL proteins predicted with STRING with a threshold of 0.7 for score. Rows shaded in grey correspond to circadian-related proteins studied here; in blue are highlighted putative circadian proteins with secondary functions.

Protein	Interactors	Interactor description	Score
BMAL1A	CRY3	Cryptochome circadian clock 3	0.993
	CRY1AA	Cryptochome circadian clock 1AA	0.988
	PER3	Period circadian clock 3	0.988
	CRY1BA	Cryptochome circadian clock 1BA	0.984
	PER2	Period circadian clock 2	0.976
	PER1B	Period circadian clock 1B	0.975
	NFIL3-5	nuclear factor, interleukin 3-regulates, member 5	0.972
	CLOCKA	Clock circadian clock regulator a	0.961
	CRY2	Cryptochome circadian clock 2	0.948
	CRY5	Cryptochrome 5	0.936
BMAL1B	CLOCKA	Clock circadian clock regulator a	0.954
	CRY1AB	Cryptochome circadian clock 1AB	0.939
	CRY1BA	Cryptochome circadian clock 1BA	0.906
	PER3	Period circadian clock 3	0.900
	CRY1AA	Cryptochome circadian clock 1AA	0.898
	CRY5	Cryptochrome 5	0.897
	BMAL1A	aryl hydrocarbon receptor nuclear translocator-like 1a	0.879
	CRY1BB	Cryptochome circadian clock 1BB	0.869
	PER1B	Period circadian clock 1B	0.861
BMAL2	CRY3	Cryptochome circadian clock 3	0.993
	CRY1AA	Cryptochome circadian clock 1AA	0.988
	PER3	Period circadian clock 3	0.988
	CRY1BA	Cryptochome circadian clock 1BA	0.984
	PER2	Period circadian clock 2	0.976
	PER1B	Period circadian clock 1B	0.975
	NFIL3-5	nuclear factor, interleukin 3-regulates, member 5	0.972
	CLOCKA	Clock circadian clock regulator a	0.961
	CRY2	Cryptochome circadian clock 2	0.948
	CRY5	Cryptochrome 5	0.936

Table S20. Summary of the results obtained by FEL analysis for pervasive negative selection in coding gene of TIMELESS protein. A threshold of 0.1 was assumed for significance level.

Protein	**Number of sites under negative selection**	**Sites under negative selection**	**p-value**
TIMELESS	3	A442	0.015
		E661	0.024
		I1231	0.005

Table S21. Analysis of patters of post-translational modifications for TIMELESS protein. Rows shaded in grey correspond to PTMs with modifications in studied populations.

		Species				
Protein	**Post-translational modification/Domain**	*S. torgalensis*	*S. aradensis*	*S. carolitertii*	*S. pyrenaicus (Tagus)*	*S. pyrenaicus (Almargem)*
TIMELESS	*Protein kinase C phosphorylation site*	20	20	19	19	19
	Casein kinase II phosphorylation site	22	23	22	22	23
	N-glycosylation site	6	6	6	6	6
	cAMP- and cGMP-dependent protein kinase phosphorylation site	6	6	6	6	6
	Tyrosine kinase phosphorylation site	1	1	1	1	1
	Amidation site	1	1	1	1	1
	Microbodies C-terminal targeting signal	1	1	1	1	1

Table S22. Patterns of protein-protein interactions for TIMELESS protein predicted with STRING with a threshold of 0.7 for score. Rows shaded in grey correspond to circadian-related proteins studied here; in green are highlighted proteins related with cell cycle regulation.

Protein	Interactors	Interactor description	Score
TIMELESS	TIPIN	Timeless-interacting protein	0.990
	MCM2	minichromosome maintenance complex component 2 (DNA helicase)	0.953
	CRY2	Cryptochome circadian clock 2	0.952
	MCM9	minichromosome maintenance complex component 9 (DNA helicase)	0.933
	POLA1	Polymerase (DNA directed), alpha 1	0.930
	MCM4	minichromosome maintenance complex component 4 (DNA helicase)	0.930
	CRY1AA	Cryptochome circadian clock 1AA	0.927
	MCM3L	minichromosome maintenance complex component 3L (DNA helicase)	0.925
	CHEK1	Checkpoint kinase 1	0.914

Figures

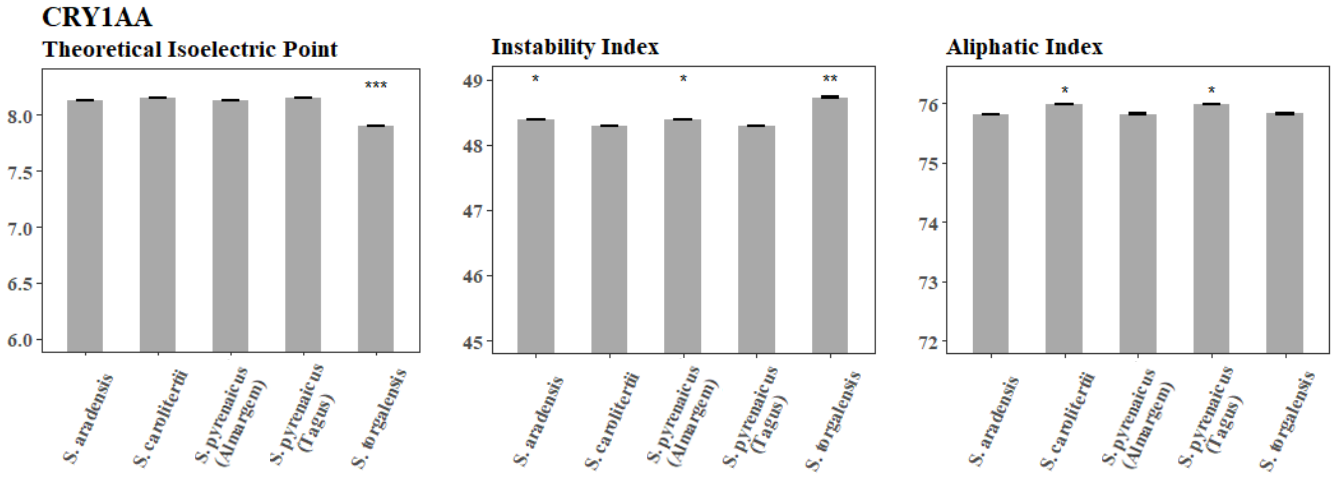

Figure S1. Predicted CRY1AA physicochemical parameters. Each bar represents the mean values for each parameter of each population (n=5) and error bars represent standard error. * p<0.05; ** p<0.01; *** p<0.001.

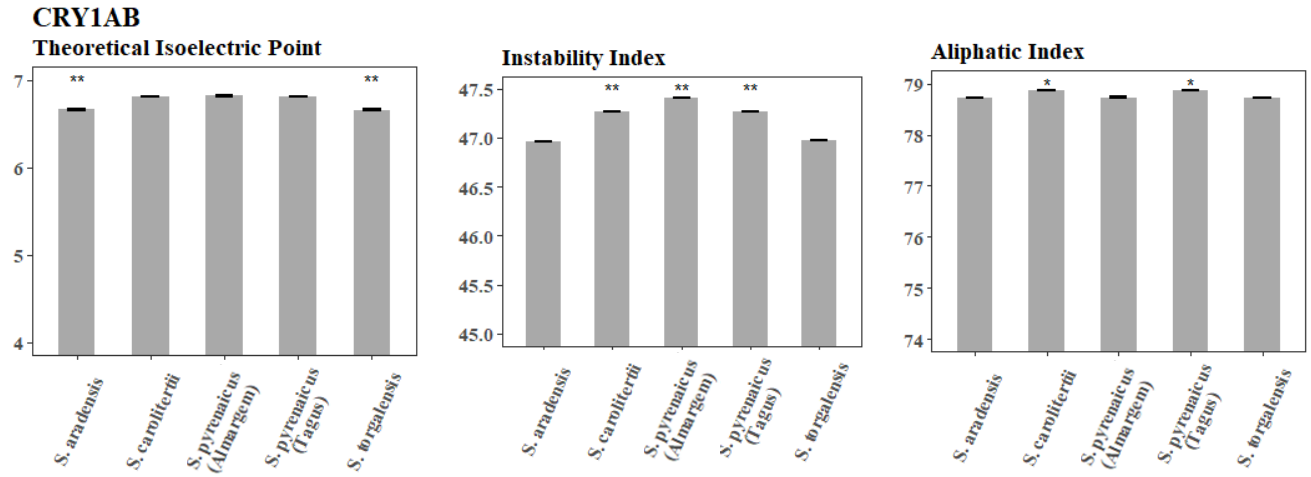

Figure S2. Predicted CRY1AB physicochemical parameters. Each bar represents the mean values for each parameter of each population (n=5) and error bars represent standard error. * p<0.05; ** p<0.01; *** p<0.001.

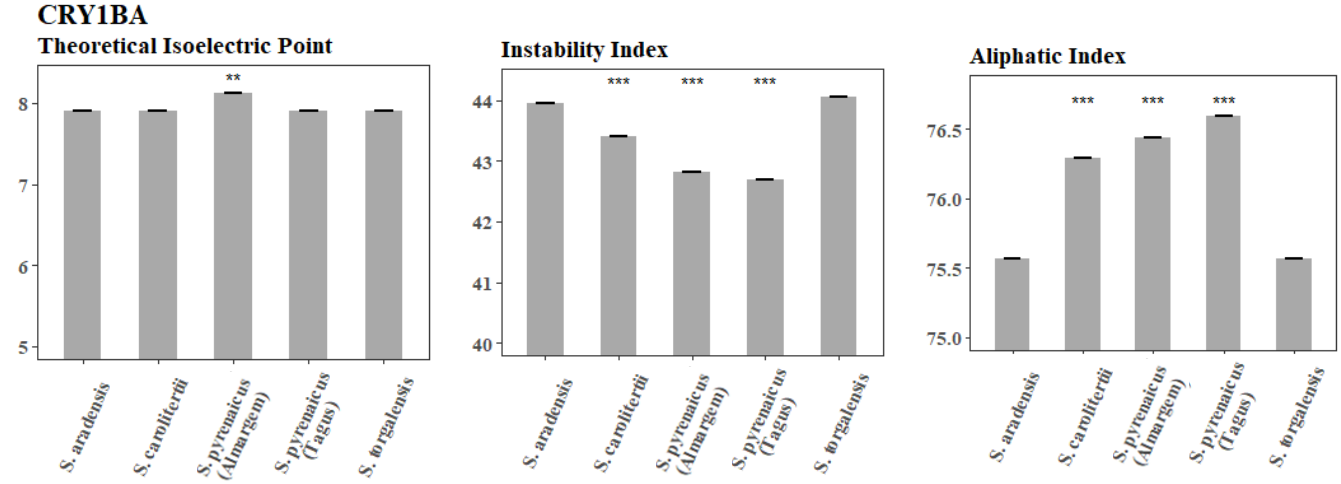

Figure S3. Predicted CRY1BA physicochemical parameters. Each bar represents the mean values for each parameter of each population (n=5) and error bars represent standard error. * p<0.05; ** p<0.01; *** p<0.001.

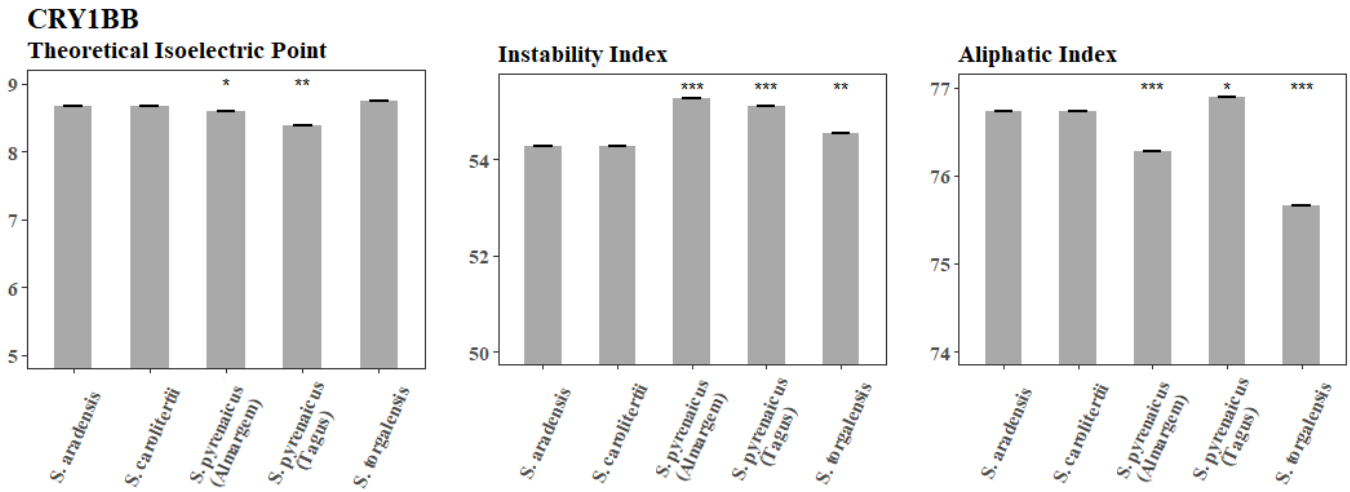

Figure S4. Predicted CRY1BB physicochemical parameters. Each bar represents the mean values for each parameter of each population (n=5) and error bars represent standard error. * p<0.05; ** p<0.01; *** p<0.001.

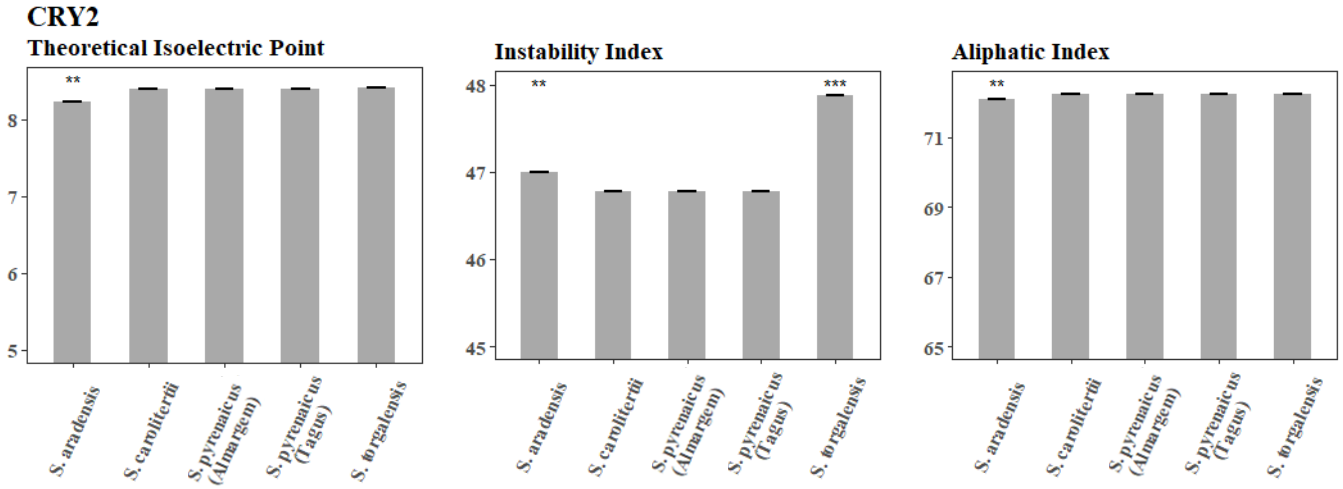

Figure S5. Predicted CRY2 physicochemical parameters. Each bar represents the mean values for each parameter of each population (n=5) and error bars represent standard error. * p<0.05; ** p<0.01; *** p<0.001.

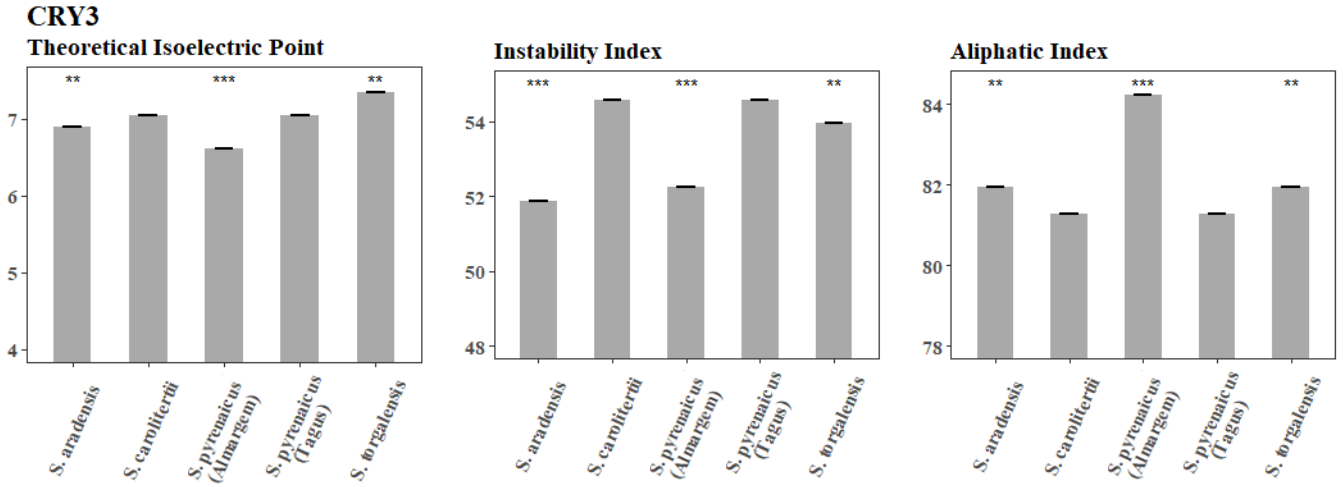

Figure S6. Predicted CRY3 physicochemical parameters. Each bar represents the mean values for each parameter of each population (n=5) and error bars represent standard error. * p<0.05; ** p<0.01; *** p<0.001.

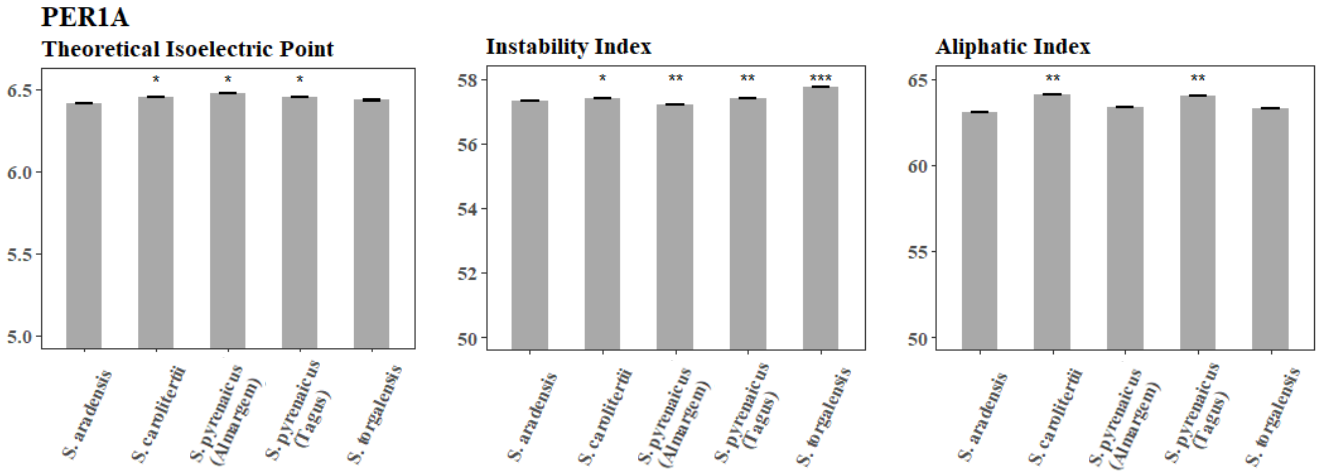

Figure S7. Predicted PER1A physicochemical parameters. Each bar represents the mean values for each parameter of each population (n=5) and error bars represent standard error. * p<0.05; ** p<0.01; *** p<0.001.

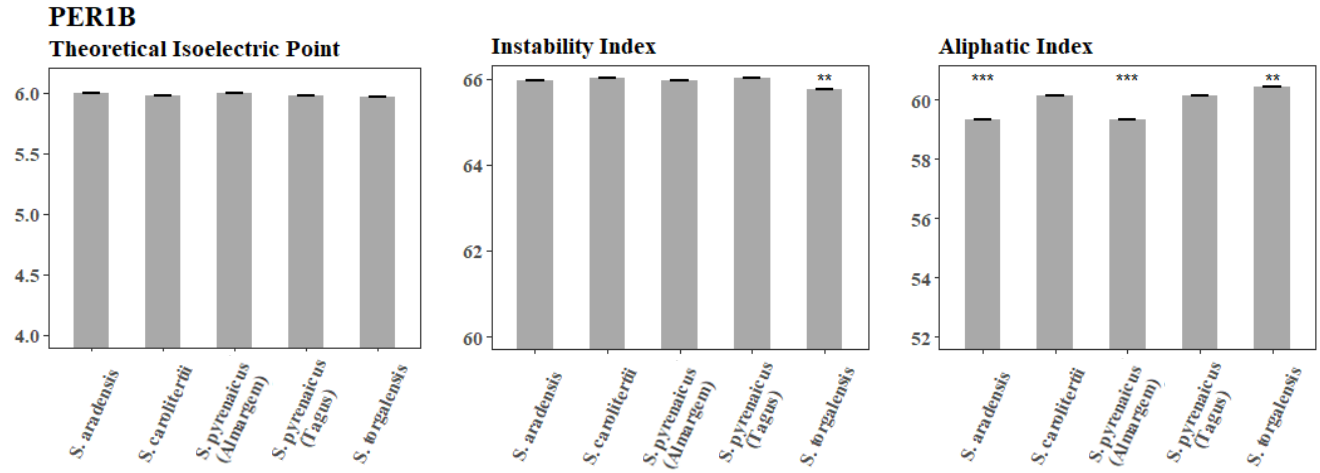

Figure S8. Predicted PER1B physicochemical parameters. Each bar represents the mean values for each parameter of each population (n=5) and error bars represent standard error. * p<0.05; ** p<0.01; *** p<0.001.

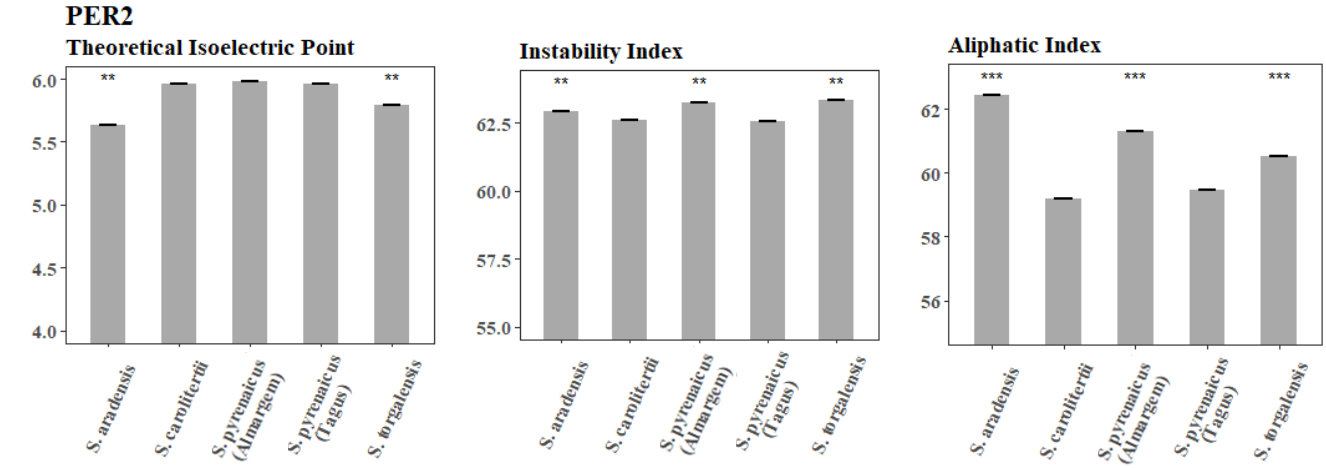

Figure S9. Predicted PER2 physicochemical parameters. Each bar represents the mean values for each parameter of each population (n=5) and error bars represent standard error. * p<0.05; ** p<0.01; *** p<0.001.

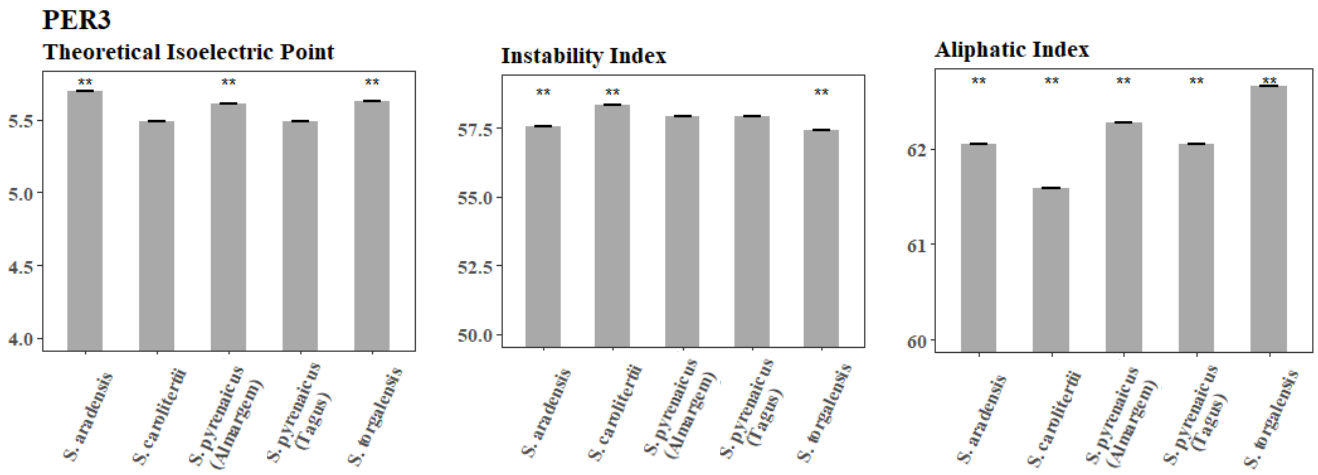

Figure S10. Predicted PER3 physicochemical parameters. Each bar represents the mean values for each parameter of each population (n=5) and error bars represent standard error. * p<0.05; ** p<0.01; *** p<0.001.

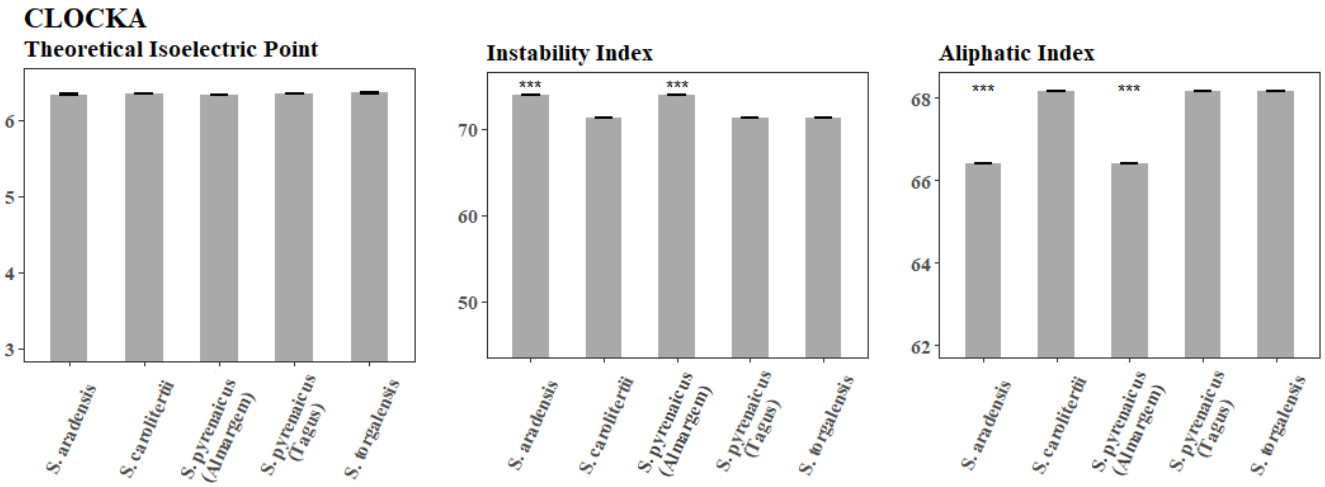

Figure S11. Predicted CLOCKA physicochemical parameters. Each bar represents the mean values for each parameter of each population (n=5) and error bars represent standard error. * p<0.05; ** p<0.01; *** p<0.001.

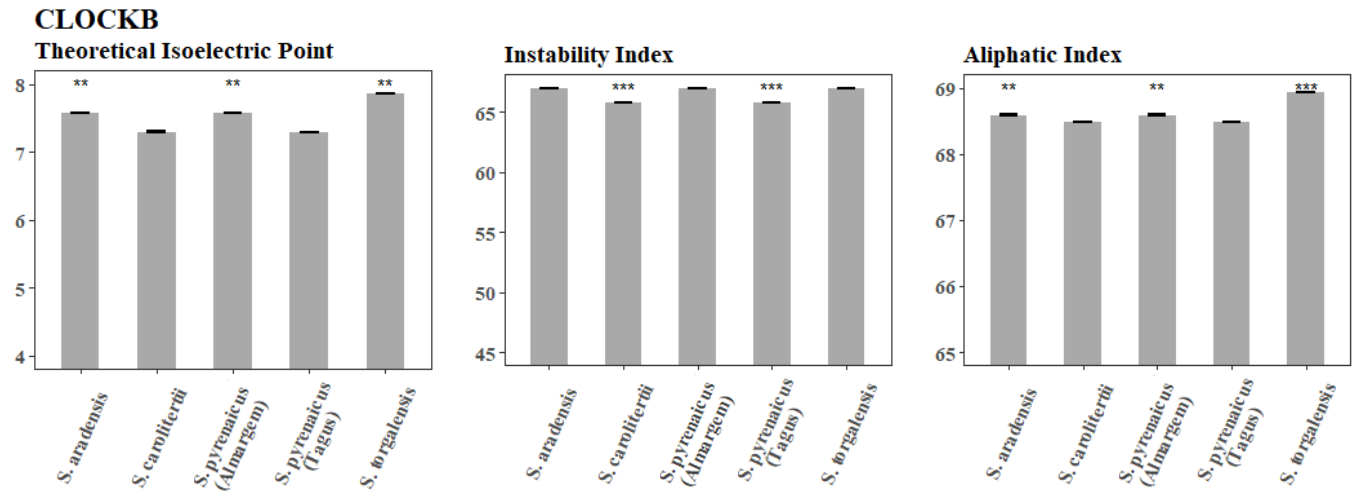

Figure S12. Predicted CLOCKB physicochemical parameters. Each bar represents the mean values for each parameter of each population (n=5) and error bars represent standard error. * p<0.05; ** p<0.01; *** p<0.001.

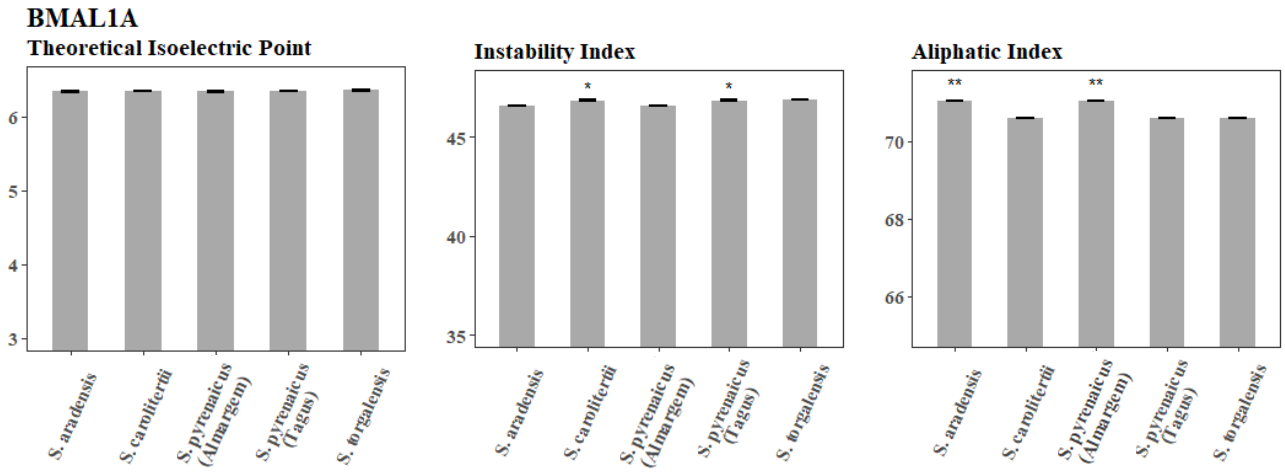

Figure S13. Predicted BMAL1A physicochemical parameters. Each bar represents the mean values for each parameter of each population (n=5) and error bars represent standard error. * p<0.05; ** p<0.01; *** p<0.001.

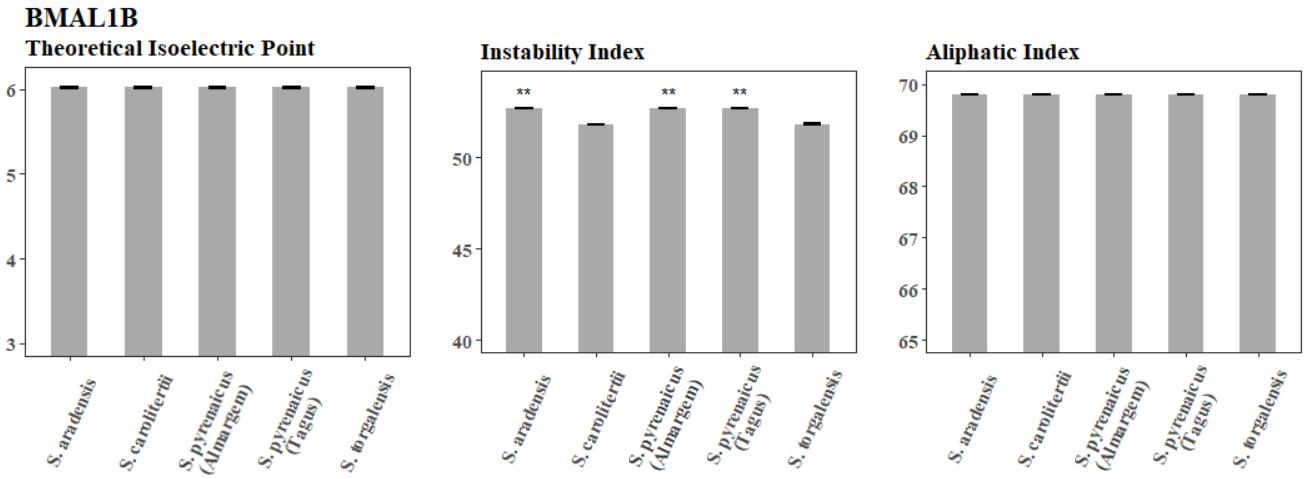

Figure S14. Predicted BMAL1B physicochemical parameters. Each bar represents the mean values for each parameter of each population (n=5) and error bars represent standard error. * p<0.05; ** p<0.01; *** p<0.001.

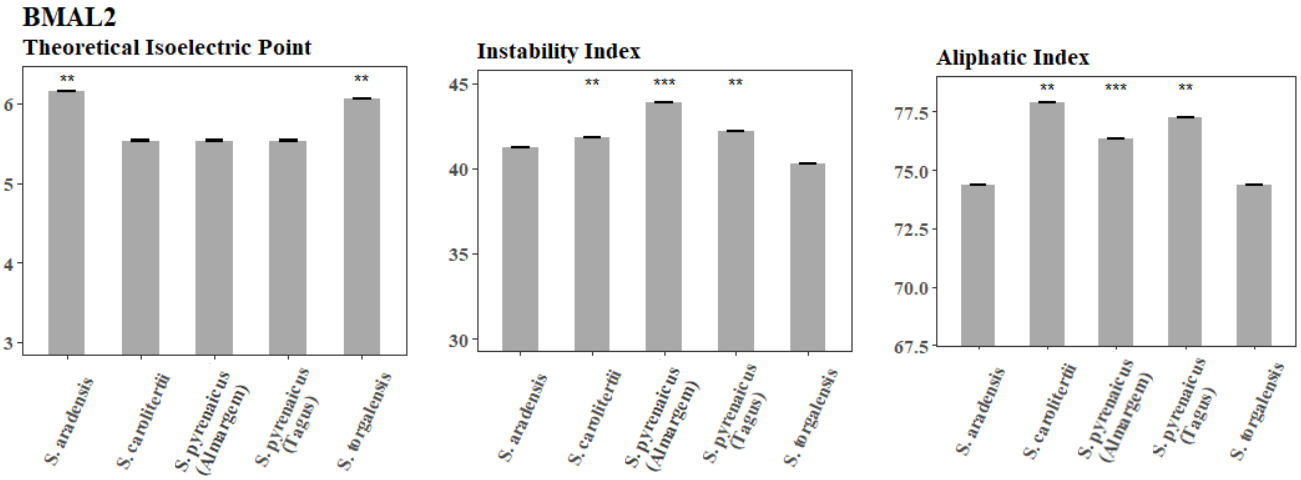

Figure S15. Predicted BMAL2 physicochemical parameters. Each bar represents the mean values for each parameter of each population (n=5) and error bars represent standard error. * p<0.05; ** p<0.01; *** p<0.001.

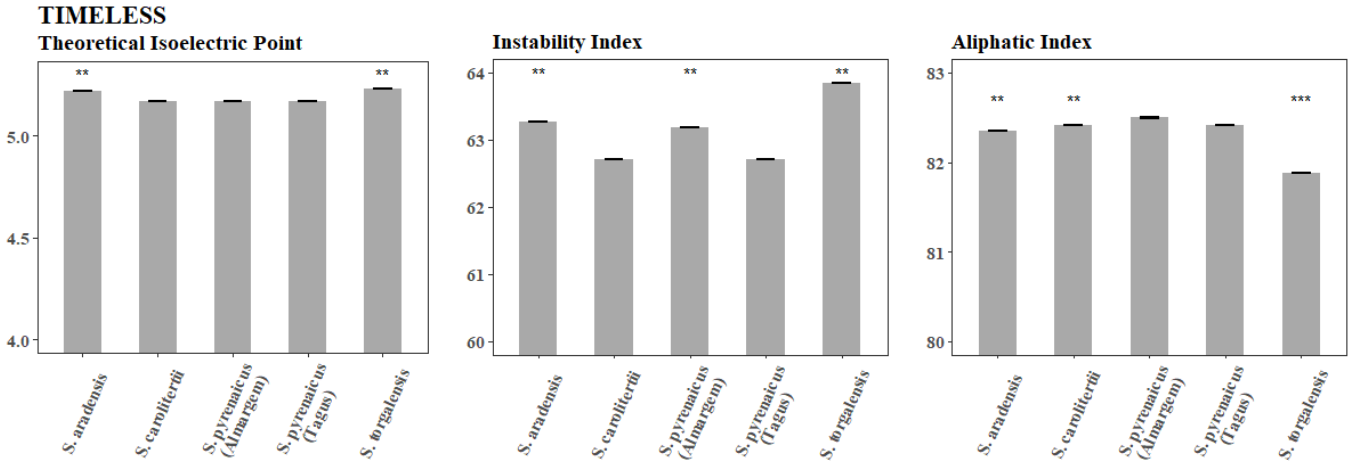

Figure S16. Predicted TIMELESS physicochemical parameters. Each bar represents the mean values for each parameter of each population (n=5) and error bars represent standard error. * $p<0.05$; ** $p<0.01$; *** $p<0.001$.